

INTRODUCTION TO PROBABILITY AND MATHEMATICAL STATISTICS



Elliot Gangaram
Queens College

Introduction to Probability and Mathematical Statistics

This text is disseminated via the Open Education Resource (OER) LibreTexts Project (<https://LibreTexts.org>) and like the thousands of other texts available within this powerful platform, it is freely available for reading, printing, and "consuming."

The LibreTexts mission is to bring together students, faculty, and scholars in a collaborative effort to provide an accessible, and comprehensive platform that empowers our community to develop, curate, adapt, and adopt openly licensed resources and technologies; through these efforts we can reduce the financial burden born from traditional educational resource costs, ensuring education is more accessible for students and communities worldwide.

Most, but not all, pages in the library have licenses that may allow individuals to make changes, save, and print this book. Carefully consult the applicable license(s) before pursuing such effects. Instructors can adopt existing LibreTexts texts or Remix them to quickly build course-specific resources to meet the needs of their students. Unlike traditional textbooks, LibreTexts' web based origins allow powerful integration of advanced features and new technologies to support learning.



LibreTexts is the adaptable, user-friendly non-profit open education resource platform that educators trust for creating, customizing, and sharing accessible, interactive textbooks, adaptive homework, and ancillary materials. We collaborate with individuals and organizations to champion open education initiatives, support institutional publishing programs, drive curriculum development projects, and more.

The LibreTexts libraries are Powered by [NICE CXone Expert](#) and was supported by the Department of Education Open Textbook Pilot Project, the California Education Learning Lab, the UC Davis Office of the Provost, the UC Davis Library, the California State University Affordable Learning Solutions Program, and Merlot. This material is based upon work supported by the National Science Foundation under Grant No. 1246120, 1525057, and 1413739.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation nor the US Department of Education.

Have questions or comments? For information about adoptions or adaptations contact info@LibreTexts.org or visit our main website at <https://LibreTexts.org>.

This text was compiled on 06/10/2025

TABLE OF CONTENTS

Licensing

Chapter 1: Week 1

- 1.1: Basic Set Theory
- 1.2: The Three Axioms of Probability
- 1.3: Basic Propositions in Probability

Chapter 2: Week 2

- 2.1: Applications
- 2.2: Introduction to Counting
- 2.3: Permutations
 - 2.3.1: The Multinomial Part 1

Chapter 3: Week 3

- 3.1: Combinations
- 3.2: The Multinomial
 - 3.2.1: The Multinomial Part 2

Chapter 4: Week 4

- 4.1: Simple Sample Spaces
- 4.2: Overcounting

Chapter 5: Week 5

- 5.1: Review 1
- 5.2: Exam 1

Chapter 6: Week 6

- 6.1: Conditional Probabilities
- 6.2: Independence

Chapter 7: Week 7

- 7.1: Bayes' Theorem

Chapter 8: Week 8

- 8.1: Discrete Random Variables
- 8.2: The Cumulative Distribution Function
- 8.3: Expected Values

Chapter 9: Week 9

- 9.1: Variances
- 9.2: The Uniform Random Variable
- 9.3: The Bernoulli and Binomial Random Variables
- 9.4: The Geometric and Negative Binomial Random Variables

- [9.5: The Poisson Random Variable](#)

Chapter 10: Week 10

- [10.1: Review 2](#)
- [10.2: Exam 2](#)

Chapter 11: Week 11

- [11.1: Continuous Random Variables](#)
- [11.2: The Uniform Continuous Distribution](#)
- [11.3: The Exponential Distribution](#)
- [11.4: The Normal Distribution](#)

Chapter 12: Week 12

- [12.1: The Central Limit Theorem](#)

Chapter 13: Week 13

- [13.1: Inferential Statistics](#)
- [13.2: Confidence Intervals](#)

Chapter 14: Week 14

- [14.1: Procedures of Hypotheses Testing and the Scientific Method](#)
- [14.2: Formulate General Research Questions](#)
- [14.3: Design Research Hypotheses and Experiment](#)
- [14.4: Collect and Analyze Experimental Data](#)
- [14.5: Report Conclusions in Non-statistical Language](#)
- [14.6: Test of Mean vs. Hypothesized Value – A Complete Example](#)
- [14.7: Type II Error and Statistical Power](#)
- [14.8: New Models for One Population Inference, Similar Procedures](#)
- [14.9: The p-value- Misconceptions and Proper Usage](#)

Chapter 15: Week 15

- [15.1: Review 3](#)
- [15.2: Final Exam](#)

[Index](#)

[Glossary](#)

[Detailed Licensing](#)

[Detailed Licensing](#)

Licensing

A detailed breakdown of this resource's licensing can be found in [Back Matter/Detailed Licensing](#).

CHAPTER OVERVIEW

Chapter 1: Week 1

[1.1: Basic Set Theory](#)

[1.2: The Three Axioms of Probability](#)

[1.3: Basic Propositions in Probability](#)

[Chapter 1: Week 1](#) is shared under a [CC BY](#) license and was authored, remixed, and/or curated by LibreTexts.

1.1: Basic Set Theory

The goal of this chapter is to understand *what is meant when we use the word probability?*

There is an informal way and a formal way to answer this. The informal way appeals towards our intuition and so allow us to first discuss the informal solution. To do so, we first make a preliminary definition.

Definition: Relative Frequency

The relative frequency of an event refers to the ratio of the number of times for which an event occurs to the the total number of trials.

For instance, suppose we were flipping a fair coin 500 times and we were interested in the relative frequency of obtaining a heads. The following code in R simulates the experiment:

Click the "run" button above to simulate the experiment. We may repeatedly click the "run" button to obtain different simulations of tossing 500 coins.

For concreteness, let us run the experiment and determine the output:

✓ Example 1.1.1



Login with LibreOne to run this code cell interactively.

If you have already signed in, please refresh the page.

Login

```
coin <- c('heads', 'tails')
flips <- sample(coin, size = 500, replace = TRUE)
freqs <- table(flips)
freqs
```

```
flips
heads tails
  255   245
```

We see that in the above simulation, we ended up with 255 heads after a total of 500 flips or trials. Hence the relative frequency is $\frac{255}{500}$. With this definition in mind, we can begin to informally understand probability.

Definition: Informal Definition of Probability

The probability of an event refers to the likelihood or relative frequency for which the event occurs if the experiment were repeated a large number of times under similar conditions.

What this definition is saying is that the probability of obtaining a head when a fair coin is tossed is 0.5 because the relative frequency of heads should be approximately 0.5 when the coin is tossed a large number of times under similar conditions. To see this, allow us to run the experiment a large number of times. For instance, let us run the experiment 500 times, 1000 times, 10,000 times and then 1,000,000 times and study the behavior of the relative frequency of heads.

As we repeat the experiment a large number of times under similar conditions, it appears that the relative frequency gets closer and close to 0.5. Thus, we may be inclined to say that the probability we obtain a heads is 0.5.

However, there are problems with this informal approach! What are these problems? Let us look back at our provisional definition of probability:

Definition: Informal Definition of Probability

The probability of an **event** refers to the likelihood or relative frequency for which the event occurs if the experiment were **repeated a large number** of times under **similar conditions**.

Problems

1. What do we mean by the term "**event**"?
2. The second problem we face is what if it is not easy to repeat this mysterious event?
3. Additionally, how large is large? In the context of the above example, someone may say that 1,000,000 tosses is sufficient. And hence the probability that a toss of a fair coin resulting in heads is given by $\frac{500008}{1000000} \neq 0.5$.
4. What do we mean by "**similar conditions**"? Let us suppose we flipped the coin once and obtained heads. If we were to replicate the coin toss under *exact conditions* then what would happen? For instance, let us suppose that we kept all variables the same. That is, we used the same energy on the coin flip, we used the same trajectory of the initial coin flip, the same starting position of the coin, and so forth. If we keep everything identical then every flip will land on heads which takes away from the inherent randomness of the coin!

We see that our informal definition of probability has some holes in it and this is problematic! In order to study probability, we first must agree what probability is. The formal definition of probability which we all agree to will be presented in the next section. This definition is rooted in the language of sets and thus we must discuss basic set theory.

We first start off with a definition of one of the most fundamental objects in mathematics, a set. Quite simply,

Definition: Set

A **set**, typically denoted by a capital letter such as A , is an unordered collection of well-defined objects. We call these objects the elements of A . If an object x is an element of A , then we write $x \in A$. If x is not an element of A , then we write that $x \notin A$.

Before we take a look at some examples, an explanation is in order about the word "well-defined" means. For our purposes, we can think of well-defined as meaning that given any object, it is clear if that object belongs to the set or not. Here are a few examples of sets:

Example of Sets 1.1.2

1. $A = \{1, 2, 3, 4, 5, 6\}$.
2. $E = \{x \mid x \text{ is a student enrolled in our Math 241 course}\}$.
3. $F = \{x \in \mathbb{Z} \mid x \leq 10\}$.
4. $G = \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}$

✓ Non-example of a Set 1.1.3

The following is *not* an example of a set:

$$A = \{x \mid x \text{ is a good drink}\} \quad (1.1.1)$$

Can you explain why?

Sets will appear in every problem that we will ever consider in this course and so it is important to have an excellent command on sets and the operations involved with them. In order to understand one of the most important sets that we will discuss, we first must make a definition.

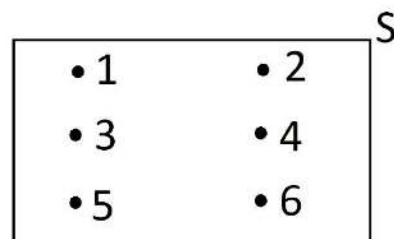
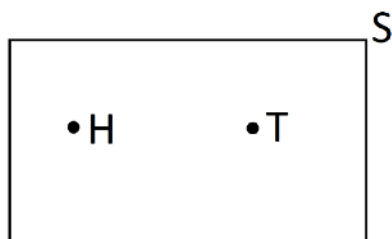
Definition: Experiment

An **experiment** is any process, real or fake, where the possible outcomes of the process can be identified prior to the completion of the process. The set of all possible outcomes from an experiment is referred to as the **sample space** and is typically denoted by S or by the Greek letter, capital omega, Ω .

✓ Examples of Experiments and Sample Spaces 1.1.4

1. The act of flipping a coin can be regarded as an experiment because before we flip the coin, we know all of the possible outcomes - we either obtain a heads or a tails. In this example, we have $S = \{H, T\}$ where the H stands for Heads and the T stands for tails.
2. Rolling a regular six-sided die is an experiment where $S = \{1, 2, 3, 4, 5, 6\}$.
3. Drawing a single card from a regular deck of 52 cards is also an experiment. Why?
4. Measuring the amount of time in seconds it takes for your phone's battery life is also an experiment. What do you think the sample space is for this experiment?

When it comes to probability, having a visual aide is helpful. For that reason, we typically sketch out the sample space of a problem. Often, we will sketch out the sample space by a rectangular box labeled S and we will list out the elements in S . For instance, the sample space mentioned in Example 1 and 2 above can be depicted as this:



Informally speaking, we will often study a set contained within a larger set which. We call this set which is contained in the larger set a **subset**.

Definition: Subset

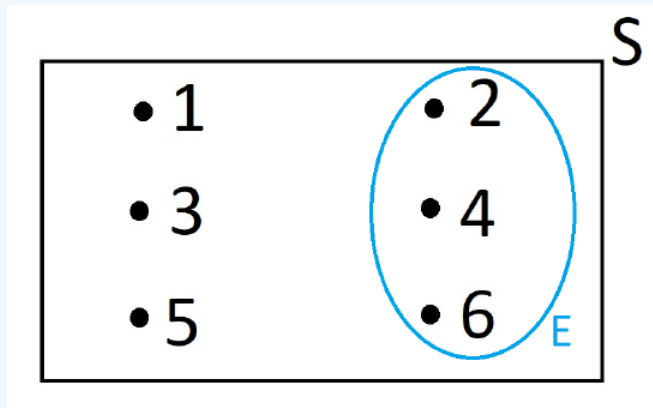
Given two sets, A and B , we say A is a **subset** of B and write $A \subseteq B$ provided that every element of A is an element of B . That is,

$$A \subseteq B \text{ if and only if } x \in A \implies x \in B \quad (1.1.2)$$

✓ Example of a subset 1.1.5

If $E = \{1, 2\}$ and $S = \{1, 2, 3, 4, 5, 6\}$ then $E \subseteq S$.

Visually, we have the following:



✓ Nonexample of a subset 1.1.6

If $A = \{1, 2, 3, 4, 5\}$ and $B = \{2, 3, 4, 5, 6\}$ then $A \not\subseteq B$. Can you explain why?

In many cases, we will be studying the subset of a sample space and thus we reserve a special name for this set.

✎ Definition: Event

Any subset, E , of a sample space S , is called an **event**.

✓ Example of events 1.1.7

1. If $S = \{H, T\}$ and $E = \{H\}$ then E is an event since $E \subseteq S$.
2. If $S = \{1, 2, 3, 4, 5, 6\}$ and $E = \{\text{we get a positive even number less than or equal to 6}\} = \{2, 4, 6\}$ then $E \subseteq S$.
3. Suppose an experiment is selecting a card at random from a regular deck of 52 cards. Then
 $S = \{2 \text{ of Hearts}, 2 \text{ of Diamonds}, 2 \text{ of Clubs}, 2 \text{ of Spades}, \dots, \text{Ace of Hearts}, \text{Ace of Diamonds}, \text{Ace of Clubs}, \dots, \text{Ace of Spades}\}$
 $E = \{5 \text{ of Hearts}, 5 \text{ of Diamonds}, 5 \text{ of Clubs}, 5 \text{ of Spades}\}$ then $E \subseteq S$.

For us, we will be using the words set and event interchangeably. More importantly, there is an algebra we can define on sets. What do we mean by this? Well, simply put, in the same way we can have rules for performing operations on numbers (like addition, subtraction, multiplication, and division) we can instead create rules for performing operations on sets (like taking their union, intersection, or complement).

✎ Definition: Union, Intersection, Empty Set, Disjoint/Mutually Exclusive

For any two events, E and F of a sample space S , we define $E \cup F$, read as " E **union** F " to be the event which consists of all outcomes that are in either E , or F , or in both E and F . That is,

$$E \cup F = \{x \in S : x \in E \text{ or } x \in F\} \quad (1.1.3)$$

We define $E \cap F$, read as " E **intersect** F ", to be the event which consists of all outcomes that are in both E and F . That is,

$$E \cap F = \{x \in S : x \in E \text{ and } x \in F\} \quad (1.1.4)$$

It should be noted that in the case that E and F share no common outcomes, we write $E \cap F = \emptyset$, read as " E intersect F equals the empty set", where $\emptyset$ refers to the set which contains no elements. If $E \cap F = \emptyset$, then we say that E and F are **disjoint** or **mutually exclusive** events.

✓ Examples of union and intersections 1.1.8

Find $E \cup F$ and $E \cap F$ for each of the following examples:

1. $E = \{0, 1, 2, 3, 4\}$ and $F = \{3, 4, 5, 6, 10\}$.
2. $E = \{1, 2, 3\}$ and $F = \{4, 5, 6\}$.

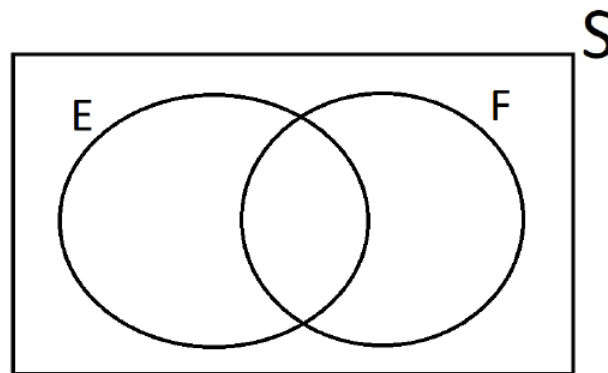
Answer

1. $E \cup F = \{0, 1, 2, 3, 4, 5, 6, 10\}$ and $E \cap F = \{3, 4\}$.
2. $E \cup F = \{1, 2, 3, 4, 5, 6\}$ and $E \cap F = \emptyset$.

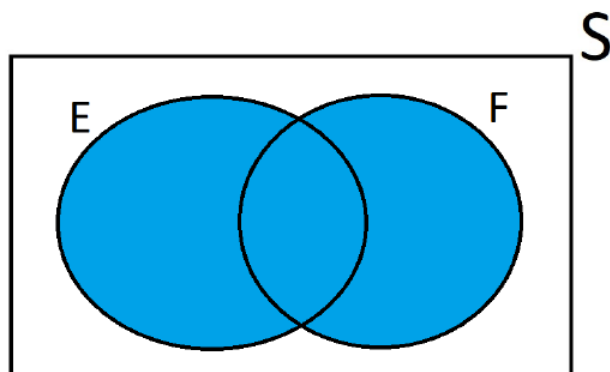
📌 Note

Informally, we can think of the set theoretic operation of union as "merging" the sets into one "larger" set. Meanwhile, the set theoretic operation of intersection only takes what is common or shared from the underlying sets.

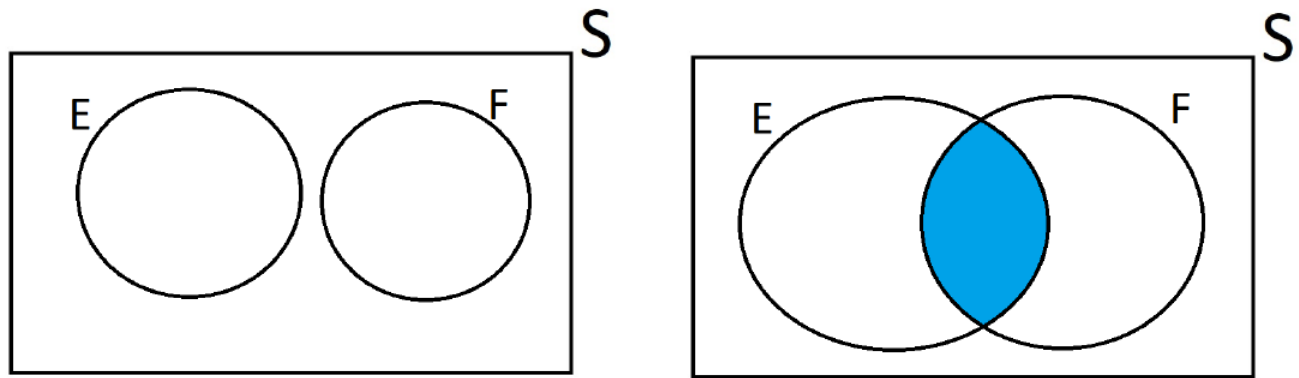
To understand these operations from a visual perspective, we can draw a Venn diagram. Here is what the Venn diagram may look like for two events:



For the above figure, can you shade the region that is referred to when we say $E \cup F$? Can you shade the region that is referred to when we say $E \cap F$?



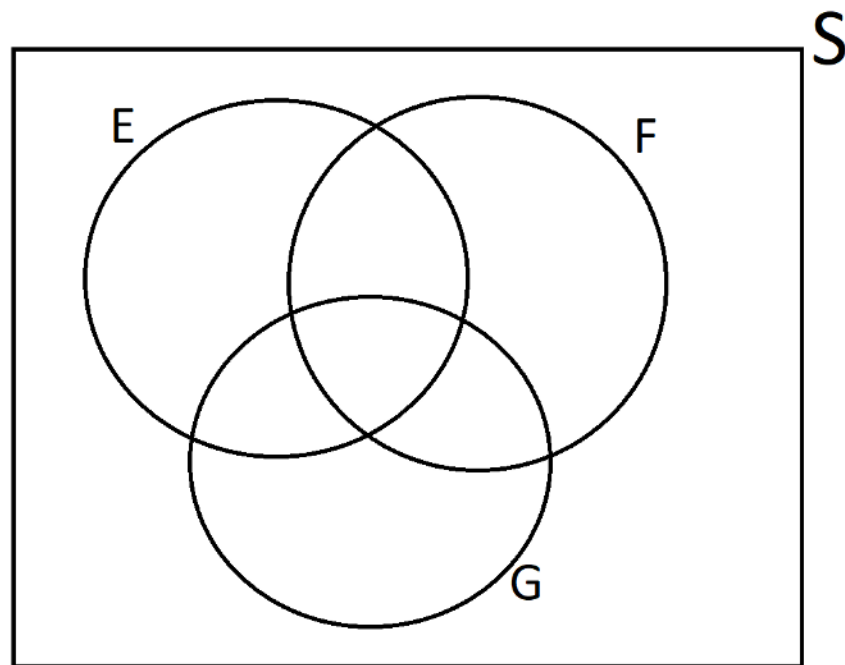
If two events are disjoint, then the Venn diagram might look like the following:



We shall now take a small jump. You see, there is no reason to restrict ourselves to studying the union or intersection of *only* two events. Instead, we may discuss the union or intersection of three events or four events or five events or k events. In fact, we can also have the union or intersection of infinitely many events. In terms of a Venn Diagram, let us see what is meant by the union of three events and the intersection of three events.

✓ Example 1.1.9

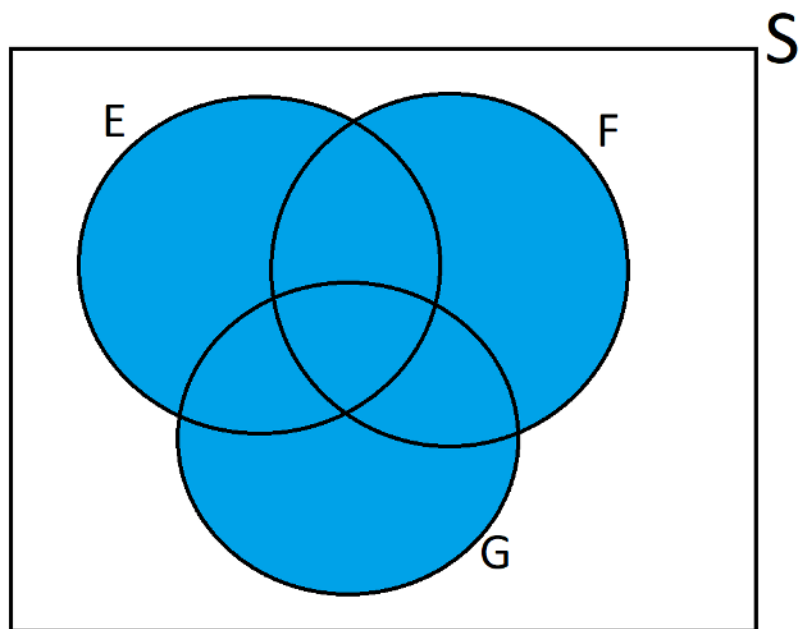
A Venn diagram for three events, E , F and G may look like the following:



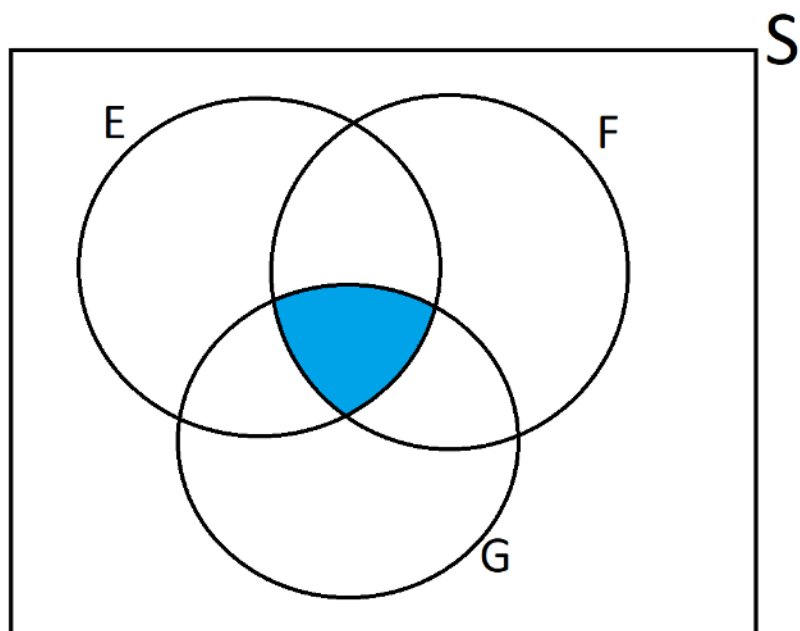
1. Shade in $E \cup F \cup G$.
2. Shade in $E \cap F \cap G$.

Answer

1.



2.



Unfortunately, we cannot draw a picture to illustrate what it means to take the union of an infinite number of sets nor can we draw the picture for what it means to take the intersection of an infinite number of sets. Why is that? Well clearly, we cannot draw an infinite number of sets! And since we do not have a concise way of doing so, we can only present it's definition.

Definition: Countably Infinite Union of Events

If $E_1, E_2, E_3, \dots$ are events in a sample space S , then the union of these events, denoted by

$$\bigcup_{i=1}^{\infty} E_i = E_1 \cup E_2 \cup E_3 \cup \dots \quad (1.1.5)$$

is defined to be the event that consist of all outcomes that are in E_n for at least one value of $n = 1, 2, 3, \dots$. That is,

$$\bigcup_{i=1}^{\infty} E_i = \{x \in S \mid x \in E_i \text{ for some } i \in \mathbb{N}\} \quad (1.1.6)$$

Definition: Countably Infinite Intersection of Events

If $E_1, E_2, E_3, \dots$ are events in a sample space S , then the intersection of these events, denoted by

$$\bigcap_{i=1}^{\infty} E_i = E_1 \cap E_2 \cap E_3 \cap \dots \quad (1.1.7)$$

is defined to be the event that consist of all outcomes that are common to all of the events E_n for $n = 1, 2, 3, \dots$.

That is,

$$\bigcap_{i=1}^{\infty} E_i = \{x \in S \mid x \in E_i \text{ for all } i \in \mathbb{N}\} \quad (1.1.8)$$

Finally, there is one last set theoretic operation we must discuss - how to take the **complement** of an event.

Definition: Complement

Given any event A of a sample space S , we define the **complement of A** , denoted by A^c or A' , to be the event containing all of the elements in S that are not in A . That is,

$$A^c = \{x \in S : x \notin A\} \quad (1.1.9)$$

✓ Example of Complement 1.1.10

Find E^c for each of the following examples:

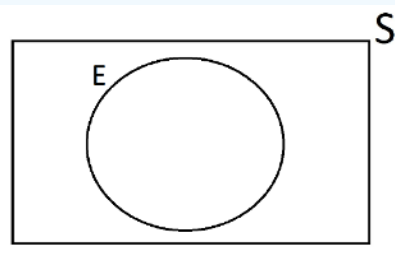
1. $S = \{0, 1, 2, 3, 4, 5, 6, 10\}$ and $E = \{0, 1, 2, 3\}$
2. $S = \{1, 2, 3, 4, 5, 6\}$ and if $E = \{2, 4, 6\}$
3. $S = \{1, 2, 3\}$ and $E = \{1, 2, 3\}$

Answer

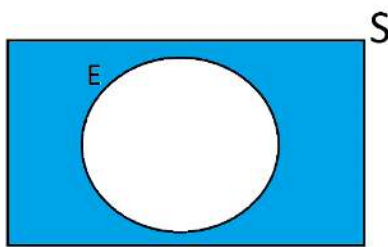
1. $E^c = \{4, 5, 6, 10\}$
2. $E^c = \{1, 3, 5\}$
3. $E^c = \emptyset$

? Exercise 1.1.11

Shade in E^c on the following Venn Diagram.



Answer



We now discuss three definitions which will play a critical part in developing our theory of probability. The first definition generalizes what it means for more than two sets to be mutually exclusive. The remaining two definitions will play a central role in Week 5 when we discuss conditional probabilities.

 Definition: Mutually Exclusive Events, Exhaustive Events, Forms a Partition on S

1. The events $E_1, E_2, \dots, E_k$ are said to be **mutually exclusive events** if $E_i \cap E_j = \emptyset$ for all $i \neq j$.
2. The events $E_1, E_2, \dots, E_k$ are said to be **exhaustive on S** if $E_1 \cup E_2 \cup \dots \cup E_k = S$.
3. The events $E_1, E_2, \dots, E_k$ are said to **form a partition on S** if $E_1, E_2, \dots, E_k$ are mutually exclusive events and exhaustive events.

 Example 1.1.12

1. If $S = \{1, 2, 3, 4, 5, 6\}$ and $E_1 = \{1, 2, 3\}$ and $E_2 = \{4, 5, 6\}$ then E_1 and E_2 are disjoint or mutually exclusive.
2. If $S = \{1, 2, 3, 4, 5, 6\}$ and $E_1 = \{1, 2, 3\}$ and $E_2 = \{4, 5, 6\}$ then E_1 and E_2 are exhaustive events.
3. If $S = \{1, 2, 3, 4, 5, 6\}$ and $E_1 = \{1, 2, 3\}$ and $E_2 = \{4, 5, 6\}$ then E_1 and E_2 form a partition on S .
4. If $S = \{1, 2, 3, 4, 5, 6\}$ and $E_1 = \{1, 2\}$ and $E_2 = \{3, 4\}$ and $E_3 = \{5, 6\}$ then E_1, E_2, E_3 forms a partition on S .
5. If $E_1 = \{1, 2, 3\}$ and $E_2 = \{4, 5, 6\}$ and $E_3 = \{7, 8, 9\}$ then E_1, E_2 and E_3 are said to be disjoint or mutually exclusive.
6. If $E_1 = \{1, 2, 3\}$ and $E_2 = \{4, 5, 6\}$ and $E_3 = \{6, 7, 8, 9\}$ then E_1, E_2 and E_3 are NOT disjoint.
7. If $S = \{1, 2, 3, 4, 5, 6\}$ and $E_1 = \{1, 2, 3, 4\}$ and $E_2 = \{4, 5, 6\}$ then E_1 and E_2 are still exhaustive events.

Finally, we end this section with a list of theorems concerning basic set theory. A good idea is to convince yourself of the truth of each statement.

 Theorem 1.1.13

$\emptyset \subseteq A$ for every set A .

Proof

Answer

$\emptyset \subseteq A$ means that $x \in \emptyset \implies x \in A$. Since the premise is false, the implication is true.

 Theorem 1.1.14

Suppose that A, B and C are subsets of a set S . Then

1. $A \subseteq A$ (the *reflexive property*).
2. If $A \subseteq B$ and $B \subseteq A$ then $A = B$ (the *anti-symmetric property*).
3. If $A \subseteq B$ and $B \subseteq C$ then $A \subseteq C$ (the *transitive property*).

 Theorem 1.1.15

If $E \subseteq F$ then $F = E \cup (E^c \cap F)$

 Theorem: Identity Laws 1.1.16

1. $A \cup \emptyset = A$
2. $A \cap S = A$

 Theorem: Idempotent Laws 1.1.17

1. $A \cup A = A$
2. $A \cap A = A$

 Theorem: Domination Laws 1.1.18

1. $A \cup \emptyset = A$
2. $A \cap \emptyset = \emptyset$

 Theorem: Complement Laws 1.1.19

1. $A \cup A^c = S$
2. $A \cap A^c = \emptyset$

 Theorem: Double Complement Law 1.1.20

$$(A^c)^c = A$$

 Theorem: Commutative Laws 1.1.21

1. $A \cup B = B \cup A$
2. $A \cap B = B \cap A$

 Theorem: Associative Laws 1.1.22

1. $A \cup (B \cap C) = (A \cup B) \cap C$
2. $A \cap (B \cup C) = (A \cap B) \cup C$

 Theorem: Distributive Laws 1.1.23

1. $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
2. $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

 Theorem: Basic DeMorgans Laws 1.1.24

1. $(A \cup B)^c = A^c \cap B^c$
2. $(A \cap B)^c = A^c \cup B^c$

 Theorem: Generalized DeMorgans Laws 1.1.25

1.
$$\left(\bigcup_{i=1}^{\infty} A_i \right)^c = \bigcap_{i=1}^{\infty} A_i^c \quad (1.1.10)$$
2.
$$\left(\bigcap_{i=1}^{\infty} A_i \right)^c = \bigcup_{i=1}^{\infty} A_i^c \quad (1.1.11)$$

1.2: The Three Axioms of Probability

In the last section, we stated that our informal definition of probability has some holes in it and this is problematic! In order to study probability, we first must agree as to what exactly a probability is. We then made a note that the formal definition of probability is rooted in the language of sets and so we studied set theory. Although we are now in a position to formally define what is a probability, there is one last topic to discuss which will hopefully grant us some insight into probability and more generally, mathematics.

In mathematics, we have a few different structures. For instance, we have definitions, theorems, axioms, lemmas, corollaries, and conjectures to name a few. Perhaps some of these terms are familiar to us but what I would like to focus on specifically are axioms and theorems. What do we mean by an axiom?

Simply put, an axiom is a starting point in mathematics. More precisely, *an axiom is a statement which we have assumed to be true*. That is, there is no proving an axiom. As an analogy, axioms are sort of our building blocks. They give us the underlying structure of the branch in mathematics we are working in and the entire branch is developed from those axioms. For us, our entire theory of probability and statistics rests upon the following three axioms:

Definition: Probability

Probability is a real-valued function P that assigns to each event A in a sample space S a number called the **probability of the event A** , denoted by $P(A)$, such that the following three properties are satisfied:

1. $P(A) \geq 0$.
2. $P(S) = 1$.
3. If $A_1, A_2, A_3, \dots$ are events and $A_i \cap A_j = \emptyset$ for all $i \neq j$ then

$$P(A_1 \cup A_2 \cup A_3 \cup \dots) = P(A_1) + P(A_2) + P(A_3) + \dots \quad (1.2.1)$$

A similar result holds for every finite collection of mutually exclusive events. That is, if $A_1, A_2, A_3, \dots, A_k$ are events and $A_i \cap A_j = \emptyset$ for all $i \neq j$ then

$$P(A_1 \cup A_2 \cup A_3 \cup \dots \cup A_k) = P(A_1) + P(A_2) + P(A_3) + \dots + P(A_k) \quad (1.2.2)$$

Simply put, here are the important aspects of the above definition in words:

Probability

1. When we say probability is a real-valued function that assigns to each event A in a sample space S a number, we mean that $P : S \rightarrow \mathbb{R}$.
2. The first axiom states that a probability is nonnegative.
3. The second axiom states that the probability of the sample space is equal to 1.
4. The third axiom states that for every collection of mutually exclusive events, the probability of their union is the sum of the individual probabilities.

Looking back at the above definition, we see that the problems we highlighted in the last section with the intuitive definition of probability are no longer present in this definition. More generally, it seems rather difficult to poke any holes in the above definition. Furthermore, we will see in the next section that the above definition of probability yields sensible results which agree with our intuition.

1. Note that the first part of Axiom 3 is what we refer to as *countable additivity* while the second part of Axiom 3 is called *countable subadditivity*. Moreover, if we assume countable additivity then we can prove that countable subadditivity holds. That is, the existence of countable subadditivity is actually a theorem. However, for simplicity, we will take both countable additivity and subadditivity as an axiom and we leave the proof that countable additivity implies countable subadditivity as an exercise.

1.2: The Three Axioms of Probability is shared under a [CC BY](#) license and was authored, remixed, and/or curated by LibreTexts.

1.3: Basic Propositions in Probability

In this section, we will develop a small list of theorems that we will use for the entire duration of the semester. For this reason, it is important that we understand the proof of each theorem, both formally and informally.

Theorem: The Probability of a Complement 1.3.1

$$P(A^c) = 1 - P(A) \quad (1.3.1)$$

Proof: Recall [The Complement Laws](#) from Section 1.1. Since $A \cup A^c = S$, it follows that

$$P(A \cup A^c) = P(S) \quad (1.3.2)$$

Again by [The Complement Laws](#), $A \cap A^c = \emptyset$ and so A and A^c are mutually exclusive. Since we have a union of mutually exclusive events, [Axiom 3](#) tells us that

$$P(A \cup A^c) = P(A) + P(A^c) \quad (1.3.3)$$

Substituting 1.3.3 into 1.3.2 yields

$$P(A) + P(A^c) = 1$$

Hence,

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

Theorem: The Probability of the Empty Set 1.3.2

$$P(\emptyset) = 0 \quad (1.3.4)$$

Proof: Recall [The Domination Laws](#) from Section 1.1. Since $A \cup \emptyset = A$, it follows that

$$P(A \cup \emptyset) = P(A) \quad (1.3.5)$$

Again by [The Domination Laws](#), $A \cap \emptyset = \emptyset$ and so A and the $\emptyset$ are mutually exclusive. Since we have a union of mutually exclusive events, [Axiom 3](#) tells us that

$$P(A \cup \emptyset) = P(A) + P(\emptyset) \quad (1.3.6)$$

Substituting 1.3.6 into 1.3.5 yields

$$P(A) + P(\emptyset) = P(A)$$

Hence,

$$P(\emptyset) = 0$$

Theorem: The Probability Concerning Subsets 1.3.3

If $E \subseteq F$ then $P(E) \leq P(F)$.

Proof: Recall from [Section 1.1](#) that if $E \subseteq F$ then $F = E \cup (E^c \cap F)$. Hence,

$$P(F) = P(E \cup (E^c \cap F)) \quad (1.3.7)$$

Notice that E and $E^c \cap F$ are mutually exclusive. Thus $E \cup (E^c \cap F)$ is a union of mutually exclusive events. Since we have a union of mutually exclusive events, [Axiom 3](#) tells us that

$$P(E \cup (E^c \cap F)) = P(E) + P(E^c \cap F) \quad (1.3.8)$$

Substitution Equation 1.3.8 into the right hand side of Equation 1.3.7 yields

$$P(F) = P(E) + P(E^c \cap F)$$

By [Axiom 1](#), we know that $P(E^c \cap F) \geq 0$ and so it follows that

$$P(E) \leq P(F)$$

Theorem: The Probability of Every Event 1.3.4

For each event A of a sample space S , $P(A) \leq 1$.

Proof: (Try this on your own before revealing the solution!)

Answer

Since $A \subseteq S$, then by the above theorem, $P(A) \leq P(S)$. By [Axiom 2](#), we know that $P(S) = 1$ and so $P(A) \leq 1$.

Our next theorem is particularly important since the result will generalize the probability of the union for *any* two events. Currently, by [Axiom 3](#), we know that if A and B are disjoint, then $P(A \cup B) = P(A) + P(B)$. However, what if A and B were *not* disjoint? In such a case, how can we find $P(A \cup B)$? The answer is given to us in the following theorem.

Theorem: The Probability of a Union of Two Events 1.3.5

For any two events, A and B of a sample space S , $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Proof:

Answer

Observe that $A \cup B = A \cup (A^c \cap B)$. Thus,

$$P(A \cup B) = P(A \cup (A^c \cap B)) \quad (1.3.9)$$

Since A and $A^c \cap B$ are mutually exclusive, then by [Axiom 3](#),

$$P(A \cup (A^c \cap B)) = P(A) + P(A^c \cap B) \quad (1.3.10)$$

Thus,

$$P(A \cup B) = P(A) + P(A^c \cap B) \quad (1.3.11)$$

Similarly, since $B = (A^c \cap B) \cup (A \cap B)$ where $(A^c \cap B)$ and $(A \cap B)$ are mutually exclusive, then by [Axiom 3](#),

$$P(B) = P(A^c \cap B) + P(A \cap B)$$

Rearranging the above equation yields

$$P(A^c \cap B) = P(B) - P(A \cap B) \quad (1.3.12)$$

Substituting Equation 1.3.12 into the right hand side of Equation 1.3.11 yields

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

Note

The above theorem establishes that $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. A visual way of understanding this is to consider the number of times each region is counted or shaded in. Consider the Venn Diagram for $A \cup B$ when A and B are not disjoint and let us study $P(A \cup B)$. Notice that there are three regions and each region is counted or shaded once. Draw it out! One region is $A \cap B^c$, the other region is $A \cap B$, and the last region is $A^c \cap B$.

Similarly, let us again consider the Venn Diagram for $A \cup B$ and let us study $P(A) + P(B) - P(A \cap B)$ by counting the number of times each region is shaded. We first shade in $P(A)$ and so there are two regions that get counted once. (Draw it out!). We then shade in $P(B)$. At this point in time the region $A \cap B^c$ is counted once, the other region is $A \cap B$ is counted *twice*, and the last region $A^c \cap B$ is counted once. In order to get this picture to match the picture described in the above paragraph, we simply subtract out $P(A \cap B)$ exactly once.

Note

Clearly the above theorem holds in the special case that A and B are disjoint. Why?

We now extend the above formula for the union of any three events.

Theorem: The Probability of a Union of Three Events 1.3.6

For any three events, E_1 , E_2 and E_3 of a sample space S ,

$$P(E_1 \cup E_2 \cup E_3) = P(E_1) + P(E_2) + P(E_3) - P(E_1 \cap E_2) - P(E_1 \cap E_3) - P(E_2 \cap E_3) + P(E_1 \cap E_2 \cap E_3) \quad (1.3.13)$$

Proof: This is left as an exercise to the reader.

Hint

We can re-write the union of these three events as the union of two events via the associative property of the union:
 $P(E_1 \cup E_2 \cup E_3) = P((E_1 \cup E_2) \cup E_3)$. We can now apply the formula for the union of two events and proceed from there.

The above theorem can also be understood by considering the Venn diagram. To end this section, we present the general inclusion-exclusion identity which can be proved by mathematical induction.

Theorem: The Probability of a Union of n Events 1.3.7

For every n events $E_1, E_2, \dots, E_n$ of a sample space S ,

$$P(E_1 \cup E_2 \cup \dots \cup E_n) = \sum_{i=1}^n P(E_i) - \sum_{i_1 < i_2} P(E_{i_1} \cap E_{i_2}) + \dots + (-1)^{k+1} \sum_{i_1 < i_2 < \dots < i_k} P(E_{i_1} \cap E_{i_2} \cap \dots \cap E_{i_k}) + \dots + (-1)^n P(E_1 \cap E_2 \cap \dots \cap E_n) \quad (1.3.14)$$

Note

Simply put, the above theorem states that the probability of the union of any n events is equal to the sum of the individual probabilities, minus the sum of the probabilities of every possible intersection of two events, plus the sum of the probabilities of every possible intersection of three events and so on.

1.3: Basic Propositions in Probability is shared under a [CC BY](https://creativecommons.org/licenses/by/4.0/) license and was authored, remixed, and/or curated by LibreTexts.

CHAPTER OVERVIEW

Chapter 2: Week 2

[2.1: Applications](#)

[2.2: Introduction to Counting](#)

[2.3: Permutations](#)

[2.3.1: The Multinomial Part 1](#)

Chapter 2: Week 2 is shared under a [CC BY](#) license and was authored, remixed, and/or curated by LibreTexts.

2.1: Applications

At this point in time, we have gone through a lengthy discussion of definitions, axioms, theorems and proofs. With these tools we have developed, we are now in a position to answer some basic questions concerning probability. However, although these questions may be basic, do not underestimate their importance! Our first example is a sample question from The Society of Actuaries' Exam P. (For those unfamiliar with actuarial science, an excellent resource is <https://www.beanactuary.org/why-actuarial-science/>. Be sure to click on the salary tab!)

Caution

Please do not be overwhelmed by the details in our answer to the next question! There are a lot of moving parts and with practice, we will be able to answer these questions quite concisely.

✓ Example 2.1.1

A survey of a group's viewing habits over the last year revealed the following information:

1. 28% watched gymnastics
2. 29% watched baseball
3. 19% watched soccer
4. 14% watched gymnastics and baseball
5. 12% watched baseball and soccer
6. 10% watched gymnastics and soccer
7. 8% watched all three sports.

Calculate the percentage of the group that watched none of the three sports during the last year.

Source: <https://www.soa.org/globalassets/assets/files/edu/edu-exam-p-sample-quest.pdf>

Solution 1

Since this is our first example, we will discuss everything in great detail. Allow us to make a few comments. When reading the above question, we notice that there is an immediate disconnect between the theory we have been learning and the way that the question is phrased. You see, at this point in time, our framework of probability is phrased in terms of sets and events. In the above question, there are no sets mentioned to us! And so the first part of this problem is to translate the word problem into a question about the probability of an event. How do we do this?

Whenever the question mentions a percentage or a probability, that is the probability of some event! And so we should introduce notation. In the above problem, let $E_G = \{\text{a person watched gymnastics}\}$, let $E_B = \{\text{a person watched baseball}\}$ and let $E_S = \{\text{a person watched soccer}\}$.

In order to translate the above percentages in terms of events, there is one last remark to make. This remark is that the English word of "and" corresponds to the set theoretic operation of intersection while the word "or" corresponds to the set theoretic operation of union. Thus, when the problem states "14% watched gymnastics and baseball" this means that the event $\{\text{watched gymnastics and baseball}\}$ can be written as $E_G \cap E_B$.

Thus we have the following:

Given	Translation
28% watched gymnastics	$P(E_G) = 0.28$
29% watched baseball	$P(E_B) = 0.29$
19% watched soccer	$P(E_S) = 0.19$
14% watched gymnastics and baseball	$P(E_G \cap E_B) = 0.14$
12% watched baseball and soccer	$P(E_B \cap E_S) = 0.12$
10% watched gymnastics and soccer	$P(E_G \cap E_S) = 0.10$
8% watched all three sports.	$P(E_G \cap E_B \cap E_S) = 0.08$

We are asked to find the percentage of the group that watched none of the three sports during the last year. Again, we must translate what the question is asking us to be in terms of set. When the question says *none of the three sports* this means *the person did NOT watch gymnastics AND the person did NOT watch baseball AND the person did NOT watch soccer*.

This means we are asked to find $P(E_G^c \cap E_B^c \cap E_S^c)$. This is now a question that we can answer with the framework we have developed. Let us study the expression $P(E_G^c \cap E_B^c \cap E_S^c)$ and see if we can relate this to any relevant theorems.

A close inspection of the expression leads us to think of the [DeMorgan's Laws](#). By the General DeMorgan's Laws, we know that

$$P(E_G^c \cap E_B^c \cap E_S^c) = P((E_G \cup E_B \cup E_S)^c)$$

We see that we have the probability of a complement and so perhaps [The Probability of a Complement Rule](#) comes to mind. Piecing everything together yields the following solution:

$$\begin{aligned} P(E_G^c \cap E_B^c \cap E_S^c) &= P((E_G \cup E_B \cup E_S)^c) \\ &= 1 - P(\underbrace{E_G \cup E_B \cup E_S}_{\text{prob. of union of events}}) \\ &= 1 - (P(E_G) + P(E_B) + P(E_S) - P(E_G \cap E_B) - P(E_B \cap E_S) - P(E_G \cap E_S) + P(E_G \cap E_B \cap E_S)) \\ &= 1 - (0.28 + 0.29 + 0.19 - 0.14 - 0.12 - 0.10 + 0.08) \\ &= 1 - 0.48 \\ &= 0.52 \end{aligned}$$

Allow us to summarize four important remarks we made in our above solution.

Remark 1

To tackle a word problem, we often have to introduce suitable notation. Whenever the question states a percentage or probability, there is some set/event being described. Thus, we should introduce some capital letter to denote that event. Example: We let $E_G = \{\text{a person watched gymnastics}\}$.

Remark 2

The word "or" corresponds to the set theoretic operation of union and the word "and" corresponds to the set theoretic operation of intersection. Example: 14% watched gymnastics and baseball means $P(E_G \cap E_B) = 0.14$.

Remark 3

In rephrasing an event, we should always ask ourselves "can we describe this event by using the words *or/and*?" Example: when the question says *none of the three sports* this means *the person did NOT watch gymnastics AND the person did NOT watch baseball AND the person did NOT watch soccer*.

Remark 4

Similar to Remark 1, translate the requested probability in terms of your notation. Example: When the question says *calculate the percentage of the group that watched none of the three sports* this means to find $P(E_G^c \cap E_B^c \cap E_S^c)$.

The solution discussed above required us to have a strong command of our theory. However, allow us to discuss the same problem from another perspective.

Example 1 (Revisited) 2.1.1

A survey of a group's viewing habits over the last year revealed the following information:

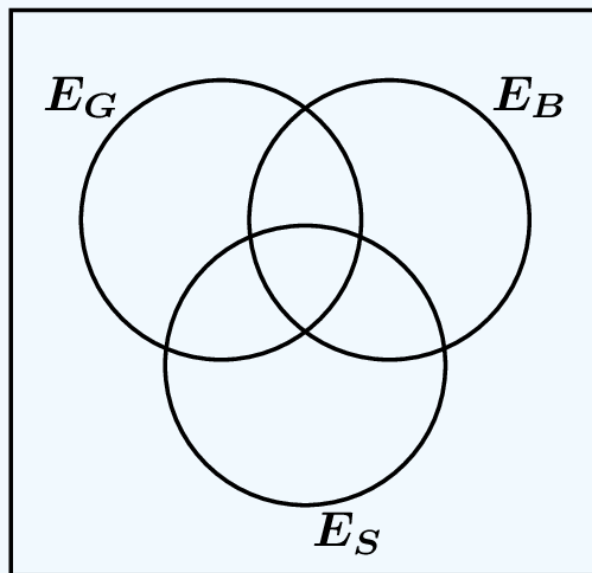
1. 28% watched gymnastics
2. 29% watched baseball
3. 19% watched soccer
4. 14% watched gymnastics and baseball

5. 12% watched baseball and soccer
6. 10% watched gymnastics and soccer
7. 8% watched all three sports.

Calculate the percentage of the group that watched none of the three sports during the last year.

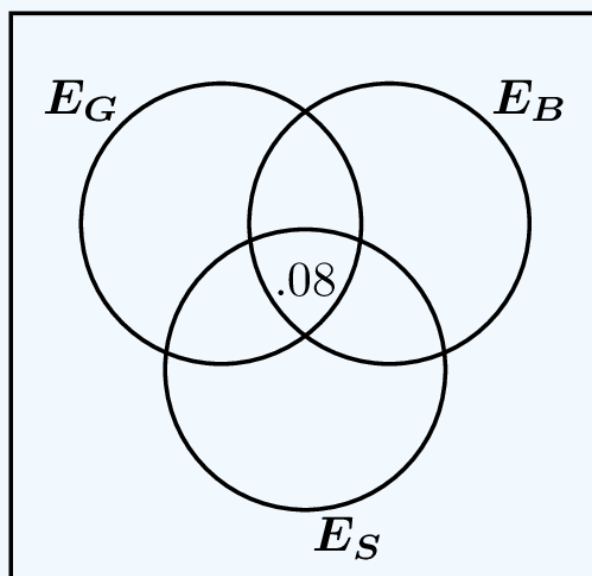
Solution 2

This time, we will study the problem via a Venn diagram:



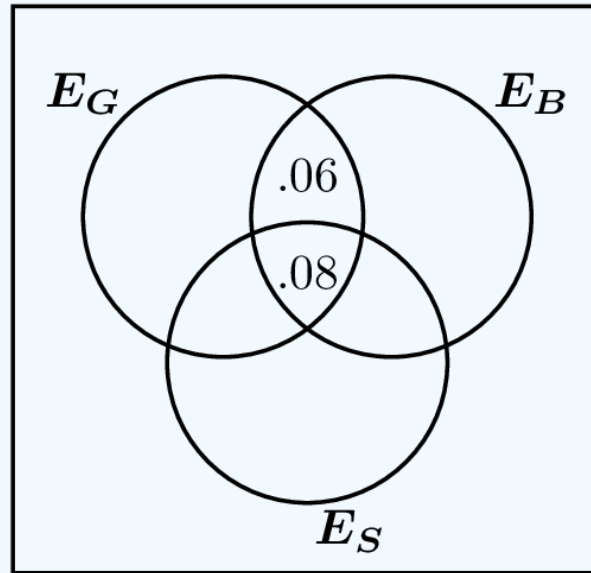
The key to working with a Venn diagram is to work *inside-out*. What does this mean?

Inside-out means we first start inside the Venn diagram, specifically the "inner most" level. That is, we first start in the intersection of all three events. According to the problem, 8% watched all three sports and so $P(E_G \cap E_B \cap E_S) = 0.08$. Thus, we have:

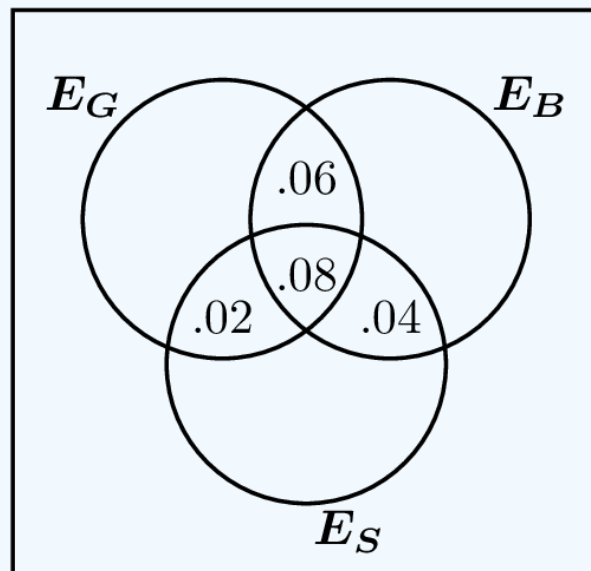


We will now work "one level" outside. According to the problem, 14% watched gymnastics and baseball meaning $P(E_G \cap E_B) = 0.14$. Since the intersection between E_G and E_B must total 14%, and we have already accounted for 8%, then

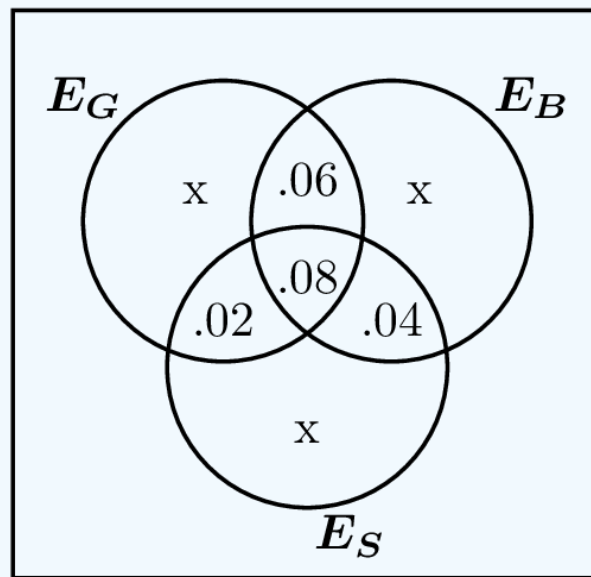
the remaining portion of the intersection must account for 6%.



By a similar logic applied to the statements "12% watched baseball and soccer" and "10% watched gymnastics and soccer", we obtain the following:

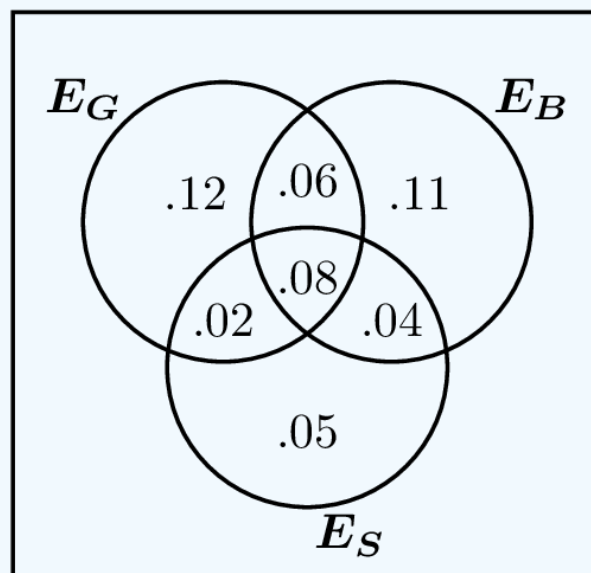


Informally speaking, we now move out another level in an attempt to find the remaining numbers in the regions that are marked off:

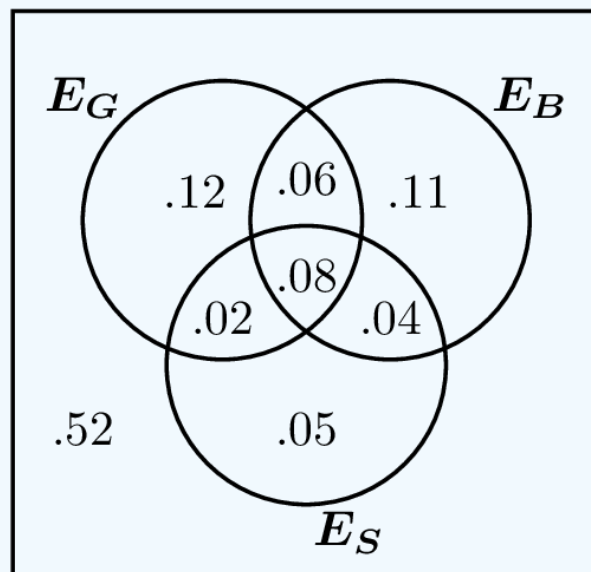


Let us first deal with the region E_G . We note that $P(E_G) = 28\%$. So far, we have accounted for $6\% + 8\% + 2\% = 16\%$. Hence the remaining portion must account for 12%.

By a similar argument concerning the regions E_B and E_S , we obtain the following:



Finally, we note that by Axiom 2, $P(S) = 1$ and so our entire diagram is to total to 100%. At this point in time, we see from the above diagram that we have accounted for 48%. Hence, 52% must be on the "outside" as shown below:



Thus, *none of the three sports* corresponds to 52% which is the same solution as above.

Here are two exercises for us to try which will test our understanding of the above concepts:

? Exercise 2.1.2

Based off your viewing habits, Netflix determines that there is a 30% chance you will like the show *Stranger Things*, a 40% chance you will like *Bridgerton* and a 25% chance you will like both shows. What is the probability that you would like at least one show? (Try to answer this question formally via the theory we discussed and then informally via a Venn diagram).

Answer

45%

? Exercise 2.1.3

According to Blockbuster,

1. 10% of viewers like horror
2. 30% of viewers like comedy
3. 5% of viewers like romance
4. 8% like horror and comedy
5. 2% like horror and romance
6. 4% like comedy and romance
7. 1% like all three genres

What is the probability that a person likes at least one of the above genres? (Try to answer this question formally via the theory we discussed and then informally via a Venn diagram).

Answer

32%

2.2: Introduction to Counting

In the last section, we saw how to use the theorems we developed to answer some questions concerning the probability of an event. In this section (and in the following), we will learn how to count. After doing so, we will then connect counting to probability. We first start off with the following definition:

Definition: Combinatorial Analysis

Combinatorial analysis refers to the mathematical theory of counting.

For many students, the topic of counting is often met with much dismay. Our goal here is to make counting as intuitive as possible. In order to do so, we will take a motivated approach to mostly every topic.

We first begin with a simple question that many students have heard of before. (However, if you did not hear of this question, then still try to answer it!).

Motivating Question: A man has 3 different shirts and 4 different pants. How many outfits does the man have? (For us, an outfit is a specification of the shirt and pants).

Most students correctly answer that the man has a total of 12 different shirts which is indeed correct. But if you are unsure as to why the answer is 12 or if you obtained a different answer, then let us explain where the 12 comes from. There are a few ways to prove your answer is correct. You may generate a chart or a tree diagram. The method of using a chart is discussed below:

✓ Example 1 2.2.1

A man has 3 different shirts and 4 different pants. How many outfits does the man have?

Solution

Let us call the three shirts S_1, S_2 and S_3 and the four pants P_1, P_2, P_3 and P_4 . We may think of an outfit as being described by an order pair (x, y) where x represents the shirt and y represents the pants. To keep track of all the outfits, we make a table as shown below:

	S_1	S_2	S_3
P_1	(S_1, P_1)	(S_2, P_1)	(S_3, P_1)
P_2	(S_1, P_2)	(S_2, P_2)	(S_3, P_2)
P_3	(S_1, P_3)	(S_2, P_3)	(S_3, P_3)
P_4	(S_1, P_4)	(S_2, P_4)	(S_3, P_4)

We now must count the number of ordered pairs which appears in our chart. We see that there are $3 \times 4 = 12$ entries, and hence 12 total possibilities.

For us, we would like to have a theorem which justifies the above multiplication. That theorem is what we will call The Basic Principle of Counting.

Theorem: The Basic Principle of Counting 2.2.2

Suppose an experiment satisfies the following two criteria:

1. The experiment is performed in two distinct parts or steps.
2. The first part of the experiment has m possible outcomes and regardless of which of these outcomes may occur, the second part of the experiment has n possible outcomes.

Then the sample space, S , has a total of $m \times n$ possible outcomes.

Proof: We use a similar technique as presented in the above solution. Let the m possible outcomes in the first part be denoted by $s_1, s_2, \dots, s_m$ and let the n possible outcomes in the second part be denoted by $w_1, w_2, \dots, w_n$. We now generate a chart to keep track of all possible outcomes:

	s_1	s_2	$\dots$	s_m
w_1	(s_1, w_1)	(s_2, w_1)	$\dots$	(s_m, w_1)
w_2	(s_1, w_2)	(s_2, w_2)	$\dots$	(s_m, w_2)
$\dots$	(s_1, w_3)	(s_2, w_3)	$\dots$	(s_m, w_3)
w_n	(s_1, w_n)	(s_2, w_n)	$\dots$	(s_m, w_n)

Counting the number of ordered pairs, we see that there are $\underbrace{m + m + \dots + m}_{n \text{ times}} = mn$ possible outcomes.

We now have a more concise way of answering Example 1.

✓ Example 1 (Revisited) 2.2.3

A man has 3 different shirts and 4 different pants. How many outfits does the man have?

Solution

We can think of choosing an outfit as a two stage experiment. We first must select a shirt and then we select a pants. Let us proceed in steps:

Step 1) We first select a shirt. There are 3 possible choices.

Step 2) We now select a pants. Regardless of the shirt selected in step 1, there are 4 possible choices.

Hence, by The Basic Principle of Counting, there are $3 \times 4 = 12$ possible outcomes.

✓ Example 2 2.2.4

Suppose there are four different routes from Ozone Park to Kew Gardens and six different routes from Kew Gardens to Flushing. How many different routes are there from Ozone Park to Flushing which passes through Kew Gardens?

Solution

We must generate a route/path from Ozone Park to Flushing which passes through Kew Gardens. To do this, we can imagine the route/path as being constructed in two steps.

Step 1) We first select a path from Ozone Park to Kew Gardens. There are 4 possible choices.

Step 2) We now select a path from Kew Gardens to Flushing. Regardless of the path selected in step 1, there are 6 possible choices.

Hence, by The Basic Principle of Counting, there are $4 \times 6 = 24$ possible paths.

There is no need to restrict ourselves to studying experiments where the experiment can be thought of as a two step process. Instead, we may wish to study experiments that have three steps or four steps or in general k steps. For instance, if we now consider an outfit to be a specification of shirt, pants, and shoes worn, then we can ask the following question: A man has 3 different shirts, 4 different pants, and 2 different shoes. How many outfits does the man have?

In situations like these, we must generalize The Basic Principle of Counting. Doing so yields The Generalized Principle of Counting.

Theorem: The Generalized Principle of Counting 2.2.5

Suppose an experiment satisfies the following two criteria:

1. The experiment is performed in k distinct parts or steps.
2. The first part of the experiment has n_1 possible outcomes and regardless of which of these outcomes may occur, the second part of the experiment has n_2 possible outcomes and regardless of which of these outcomes may occur, the third part of the experiment has n_3 possible outcomes, and $\dots$, the $(k-1)^{st}$ part of the experiment has n_{k-1} possible outcomes and regardless of which of these outcomes may occur, the k^{th} part of the experiment has n_k possible outcomes.

Then the sample space, S , has a total of $n_1 \times n_2 \times n_3 \times \dots \times n_k$ possible outcomes.

✓ Example 3 (Revisited) 2.2.6

A man has 3 different shirts, 4 different pants, and 2 different shoes. How many outfits does the man have?

Solution

We can now think of choosing an outfit as a three stage experiment. We first must select a shirt, then a pants, and then select the shoes. Let us proceed in these steps:

Step 1) We first select a shirt. There are 3 possible choices.

Step 2) We now select a pants. Regardless of the shirt selected in step 1, there are 4 possible choices.

Step 3) We now select the shoes. Regardless of the shirt and pants selected, there are 2 possible choices.

Hence, by The Generalized Principle of Counting, there are $3 \times 4 \times 2 = 24$ possible outcomes.

Although this completes the above example, it might be useful to note how The Generalized Principle of Counting can be thought of as an iterated version of The Basic Principle of Counting. To show this, allow us to consider the three stage experiment above as two separate two-stage experiments. In the first stage, we select a shirt and pants. Then in the second stage, we select a pair of shoes. Similar to what was done before, allow us to denote the three shirts by S_1, S_2 and S_3 , the four pants by P_1, P_2, P_3 and P_4 , and the two shoes by J_1 and J_2 . Doing so gives us the following chart:

	J_1	J_2
(S_1, P_1)	$((S_1, P_1), J_1)$	$((S_1, P_1), J_2)$
(S_1, P_2)	$((S_1, P_2), J_1)$	$((S_1, P_2), J_2)$
(S_1, P_3)	$((S_1, P_3), J_1)$	$((S_1, P_3), J_2)$
(S_1, P_4)	$((S_1, P_4), J_1)$	$((S_1, P_4), J_2)$
(S_2, P_1)	$((S_2, P_1), J_1)$	$((S_2, P_1), J_2)$
(S_2, P_2)	$((S_2, P_2), J_1)$	$((S_2, P_2), J_2)$
(S_2, P_3)	$((S_2, P_3), J_1)$	$((S_2, P_3), J_2)$
(S_2, P_4)	$((S_2, P_4), J_1)$	$((S_2, P_4), J_2)$
(S_3, P_1)	$((S_3, P_1), J_1)$	$((S_3, P_1), J_2)$
(S_3, P_2)	$((S_3, P_2), J_1)$	$((S_3, P_2), J_2)$
(S_3, P_3)	$((S_3, P_3), J_1)$	$((S_3, P_3), J_2)$
(S_3, P_4)	$((S_3, P_4), J_1)$	$((S_3, P_4), J_2)$

✓ Example 4: License Plates 2.2.7

How many different 7-place license plates are possible if the first three places are to be occupied with letters and the last four places are to be occupied with the digits where each digit ranges from 0-9 inclusively? You may assume that repetition of letters and digits are allowed.

Solution

A license plate may be thought of as a seven stage experiment. We proceed in steps:

1. Select a letter for position 1: 26 choices
2. Select a letter for position 2: 26 choices
3. Select a letter for position 3: 26 choices
4. Select a digit for position 4: 10 choices
5. Select a digit for position 5: 10 choices
6. Select a digit for position 6: 10 choices
7. Select a digit for position 7: 10 choices

By The Generalized Principle of Counting, there are $26^3 \times 10^4 = 175,760,000$ different license plates.

Allow us to modify the above example by considering how the solution would change if repetition was not allowed.

✓ Example 5: License Plates 2.2.8

How many different 7-place license plates are possible if the first three places are to be occupied with letters and the last four places are to be occupied with the digits where each digit ranges from 0-9 inclusively?

Solution

A license plate may be thought of as a seven stage experiment. We proceed in steps:

1. Select a letter for position 1: 26 choices
2. Select a letter for position 2: 25 choices
3. Select a letter for position 3: 24 choices
4. Select a digit for position 4: 10 choices
5. Select a digit for position 5: 9 choices
6. Select a digit for position 6: 8 choices
7. Select a digit for position 7: 7 choices

By The Generalized Principle of Counting, there are $26^3 \times 10^4 = 78,624,000$ different license plates.

At this point in time, we may feel that the above framework seems like overkill for such a simple problem and it appears to be unnecessary. However, our counting questions will become more complicated and the process will not be as short as the above problem. For those types of complex problems, we will need the above framework and so we will continue to stick to this framework to gain familiarity with it. Moreover, our focus at this point is not the end answer. Instead, our objective is to write a clear solution.

✓ Example 6 - Critique This Solution! 2.2.9

Let us suppose we have the following simplified model of Chipotle. At Chipotle, we must select:

1. One of five possible entrees (a burrito, burrito bowl, Quesadilla, salad, or taco).
2. One of seven different proteins (chicken al pastor, chicken, steak, barbacoa, carnitas, sofritas, veggie).
3. One of three possible rices (white rice, brown rice, no rice).
4. One of three possible beans (black beans, pinto beans, no beans).
5. Any amount of ten different toppings (guacamole, fresh tomato salsa, roasted chili-corn salsa, tomatillo green chili salsa, tomatillo-red chili salsa, sour cream, fajita veggies, cheese, romaine lettuce, queso blanco).

How many different dishes are available to a customer?

Critique This Solution:

By The Generalized Principle of Counting, there are $5 \times 7 \times 3 \times 3 \times 10 = 3150$ possible outcomes. Is this solution correct? What do you think?

Answer

The above solution is incorrect! Do not fall into the trap that The Basic Principle of Counting states that we simply multiple a bunch of numbers. Allow us to use our framework to fully answer this question:

1. Select one entree: 5 choices
2. Select one protein: 7 choices.
3. Select one option for rice: 3 choices
4. Select one option for beans: 3 choices
5. Select one option for guacamole: 2 choices
6. Select one option for fresh tomato salsa: 2 choices
7. Select one option for roasted chili-corn salsa: 2 choices
8. Select one option for tomatillo green chili salsa: 2 choices
9. Select one option for tomatillo red salsa: 2 choices
10. Select one option for sour cream: 2 choices
11. Select one option for fajita veggies: 2 choices
12. Select one option for cheese: 2 choices
13. Select one option for romaine lettuce: 2 choices
14. Select one option for queso blanco: 2 choices

By The Generalized Principle of Counting, there are $5 \times 7 \times 3 \times 3 \times 2^{10} = 322,560$ possible outcomes.

The above example illustrates that we should be careful while applying The Generalized Principle of Counting. Specifically, proceeding in steps can help us minimize mistakes.

2.2: Introduction to Counting is shared under a [CC BY](#) license and was authored, remixed, and/or curated by LibreTexts.

2.3: Permutations

In this section, we will discuss the theory of permutations. Before doing so, it is important to note how all of our counting structures will be related to each other. The following flowchart summarizes how each structure is naturally derived from the one before it:

This means rather than focusing on the differences between the different counting structures, we should instead focus on crafting a narrative on how each structure leads to another. In this section, we will discuss how The Generalized Principle of Counting naturally leads to the idea of a permutation.

In order to discuss the theory of permutations, we first define what a permutation is.

Definition: Permutation

Suppose a set has n distinct elements and an experiment consists of selecting k of the elements one at a time without replacement. Let each outcome consist of the k elements in the order selected. Each such outcome is called a permutation of n elements taken k at a time or is called an ordered sample without replacement.

Moreover, if the k elements are selected in the following order: $s_1, s_2, s_3, \dots, s_k$, then we denote this permutation of n elements taken k at a time by the following notation: $(s_1, s_2, s_3, \dots, s_k)$.

There is a lot to take in the above definition and so allow us to highlight the important assumptions above.

Assumptions of a Permutation

In order to have a permutation, the following assumptions must be met:

1. We have n distinct elements.
2. We are selecting k of the elements.
3. The selection takes place without replacement (meaning there is no repetition of elements).
4. For some reason, we wish to note the order of selection. That is, there has to be something special or unique to being chosen first, and something special or unique to being chosen second, and so on and so forth.

Once all four assumptions are met, we can denote the permutation by writing the elements between parenthesis and placing the elements in the appropriate order.

Here is a classic example of a scenario involving a permutation.

Example 1 2.3.1

Suppose that in a race between 26 runners, we are asked to guess who receives the gold, silver and bronze medals. Explain why our guesses can be thought of as a permutation of 26 elements taken 3 at a time. Can you write down a few different outcomes?

Solution

In order to explain why our guesses can be thought of as permutations of 26 elements taken 3 at a time, we simply verify that all four assumptions outlined above apply to the given situation.

1. We have 26 distinct runners.
2. We are selecting 3 of the runners.
3. The selection takes place without replacement since once we choose a runner to receive gold, that same runner cannot also get silver. Hence, there is no repetition.
4. The order of selection should be noted since there is something special or unique to being first (which is receiving gold), something special or unique to being second (which is receiving silver), and something special or unique to being third (which is receiving bronze).

Hence an outcome can be viewed as a permutation of 26 elements taken 3 at a time.

Now suppose the runners are labeled $A, B, C, D, \dots, Z$. Here is an example of a permutation of 26 elements taken 3 at a time: (M, A, T) . In this outcome, runner M placed first, runner A placed second, and runner T placed third.

Here is another (different!) outcome: (T, A, M) . This outcome is different because here, runner T placed first, runner A placed second, and runner M placed third.

Finally, although this was not asked in the problem, note that (Q, V, Q) is not a possible outcome since in a permutation, there is no repetition.

Intuition of a Permutation

Informally speaking, a permutation of n elements taken k at a time is just simply an ordered k -tuple without replacement. For instance, a permutation of n elements taken 2 at a time is just simply an ordered pair (x, y) . Meanwhile, a permutation of n elements taken 3 at a time is just simply an ordered triple (x, y, z) , so on and so forth.

With this idea in mind, allow to try and count the number of permutations of n elements taken k at a time. We first start off with a concrete example and then generalize.

Example 2 2.3.2

Reconsider the race between 26 runners mentioned in Example 1. If we were asked to guess the outcome on who receives gold, silver and bronze, then how many outcomes are possible?

Solution

By the previous discussion, we know that an outcome can be considered a permutation of 26 elements taken 3 at a time. In this question, we are interested in finding the total number of permutations of 26 elements taken 3 at a time. To do this, we note that an outcome is simply an ordered triple that resembles $(_, _, _)$. To find the number of such ordered triples, we can think of this as a three-stage experiment.

1. Choose a runner for position 1: 26 choices
2. Choose a runner for position 2: 25 choices
3. Choose a runner for position 3: 24 choices

By The Generalized Principle of Counting, there are $26 \times 25 \times 24 = 15,600$ different permutations.

With the above example in mind, allow us to consider the more general problem:

Example 3 2.3.3

How many permutations are there of n elements taken k at a time? (Try this on your own before checking the answer!)

Answer

In order to answer this question, we first must understand what the problem is asking us. Based off our intuition of a permutation, this problem is asking us to find the total number of ordered k -tuples without replacement. This means we must find the number of ways to fill in the following structure: $(\underbrace{\quad}_1, \underbrace{\quad}_2, \underbrace{\quad}_3, \dots, \underbrace{\quad}_{k-1}, \underbrace{\quad}_k)$.

We can think of filling in the above spaces as a k -stage experiment. We proceed in steps:

1. Choose an element for position 1: there are n choices.
2. Choose an element for position 2: there are $n - 1$ choices.
3. Choose an element for position 3: there are $n - 2$ choices.

...

...

...

k) Choose an element for position k : there are $n - (k - 1) = n - k + 1$ choices.

By The Generalized Principle of Counting, the number of permutations of n elements taken k at a time is given by $n \times (n - 1) \times (n - 2) \times \dots \times (n - k + 1)$.

The above example proves the following theorem.

Theorem: Total Number of Permutations 2.3.4

The number of permutations of n elements taken k at a time is given by $n \times (n - 1) \times (n - 2) \times \dots \times (n - k + 1)$.

Proof: Already done in Example 3.

For convenience, we often denote the expression $n \times (n - 1) \times (n - 2) \times \dots \times (n - k + 1)$ by $P_{n,k}$ which is read as "the number of permutations of n elements taken k at a time". Hence we make the following slight addition to the above theorem:

Theorem: Total Number of Permutations 2.3.5

The number of permutations of n elements taken k at a time is given by $P_{n,k} = n \times (n - 1) \times (n - 2) \times \dots \times (n - k + 1)$.

Above, we developed a formula for the number of permutations of n elements taken k at a time. Allow us to see what happens if we are asked for the number of permutations of n elements taken n at a time.

Example 4 2.3.6

How many permutations are there of n elements taken n at a time? (Try this on your own before checking the answer!)

Answer

In order to answer this question, we first must understand what the problem is asking us. Based off our intuition of a permutation, this problem is asking us to find the total number of ordered n -tuples without replacement. That is, given n distinct elements, we are asking how many ways are there to select n elements and order them. This means we must find the number of ways to fill in the following structure: $(\underbrace{}, \underbrace{}, \underbrace{}, \dots, \underbrace{\phantom{a_{n-1}}}, \underbrace{})$.

We can think of filling in the above spaces as a n -stage experiment. We proceed in steps:

1. Choose an element for position 1: there are n choices.
2. Choose an element for position 2: there are $n - 1$ choices.
3. Choose an element for position 3: there are $n - 2$ choices.

...

...

...

- n) Choose an element for position k : there is 1 choice.

By The Generalized Principle of Counting, the number of permutations of n elements taken n at a time is given by $n \times (n - 1) \times (n - 2) \times \dots \times 3 \times 2 \times 1$.

We denote the above expression as $n!$.

Formally, allow us to make the following definition and summarize the above result as a theorem.

 Definition: $n!$

Given a positive integer n , we define $n!$, read as " n factorial" to be

$$n! = n(n-1)(n-2) \dots (3)(2)(1)$$

For the special case that $n = 0$, we define $0! = 1$.

 Theorem 2.3.7

Given n distinct objects, there are $n!$ ways to arrange them.

Proof: Already done in Example 4.

✓ Example 5 2.3.8

How many ways are there to arrange 6 distinct books on a bookshelf?

Solution

By the above theorem, there are $6! = 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 720$ different arrangements.

We present one last theorem for this section followed by three examples. The following theorem establishes how the number of permutations of n elements taken k at a time can be expressed in terms of factorials.

 Theorem 2.3.1

$$P_{n,k} = \frac{n!}{(n-k)!}$$

Proof: The proof is left as an exercise to the reader.

✓ Example 6 2.3.9

Suppose that from a group of 30 people, a president, vice president, secretary and treasurer are to be chosen. How many different outcomes are possible?

Solution 1

For this question, allow us to proceed in steps.

1. We first choose a president. There are 30 choices.
2. We then choose a vice president. There are 29 choices.
3. We then choose a secretary. There are 28 choices.
4. We finally choose a treasurer. There are 27 choices.

By The Generalized Principle of Counting, we see that there are $30 \times 29 \times 28 \times 27 = 657,720$ different choices. (Notice that by invoking The Generalized Principle of Counting, we are saying that each step above is different or distinct. This will be important in the next example!)

Solution 2

We see that the above question satisfies the four assumptions required to have a permutation. That is, we can think of an outcome as an ordered 4-tuple such as $(\underbrace{\quad}_{\text{pres.}}, \underbrace{\quad}_{\text{vp.}}, \underbrace{\quad}_{\text{sec.}}, \underbrace{\quad}_{\text{treas.}})$. By our [theorem](#), the number of different permutations of 30 elements taken 4 at a time is given by $P_{30,4} = 30 \times 29 \times 28 \times 27 = 657,720$ choices.

✓ Critique This Solution! 2.3.10

Suppose that from a group of 30 people, a president, vice president, and two secretaries are to be chosen. How many different outcomes are possible?

Critique This Solution:

For this question, allow us to proceed in steps.

1. We first choose a president. There are 30 choices.
2. We then choose a vice president. There are 29 choices.
3. We then choose a secretary. There are 28 choices.
4. We then choose another secretary. There are 27 choices.

By The Generalized Principle of Counting, we see that there are $30 \times 29 \times 28 \times 27 = 657,720$ different choices.

Alternatively, we see that the above question satisfies the four assumptions required to have a permutation. That is, we can think of an outcome as an ordered 4-tuple such as $(\underbrace{\quad}_{pres.}, \underbrace{\quad}_{vp.}, \underbrace{\quad}_{sec.}, \underbrace{\quad}_{sec.})$. By our [theorem](#), the number of different permutations of 30 elements taken 4 at a time is given by $P_{30,4} = 30 \times 29 \times 28 \times 27 = 657,720$ choices. Hence we get the same answer.

Do you think the above solution is correct or incorrect? Why?

Answer

Unfortunately, the above solution is incorrect! It is incorrect to use The Generalized Principle of Counting. Recall that The Generalized Principle of Counting requires us to have distinct steps! In the proposed solution, steps 3 and 4 are the same.

Additionally, the second solution in using the permutation is incorrect because we have violated the requirement discussed under the assumptions of a permutation. To have a permutation, there must be something special about the order. Here, there is no difference between the third person being chosen and the fourth and hence we no longer have a permutation.

Thus the proposed solutions are incorrect. In order to correctly answer this problem, we need an additional counting structure that will be discussed in the next section.

✓ Example 2.3.11

In how many different ways can 12 people be seated in a row if two of the people, A and B , must be seated together?

Critique This Solution:

We have 12 distinct people and we must arrange all 12 people. By our [theorem](#) above, there are $(12! = 479,001,600)$ ways to do so.

Do you think the above solution is correct or incorrect? Why?

Answer

Unfortunately, the above solution is incorrect! While the statement itself is true, (there are indeed $12!$ ways to arrange 12 people), the statement does not answer the appropriate question. In $12!$, we are obtaining all possible arrangements of the 12 people and hence we get arrangements where person A is not seated with person B . So how can we fix this?

Here is some advice: whenever a problem requires certain elements to stick together, a good idea is to "bundle" those elements together. For instance, initially we have the following diagram:

$$\boxed{A}, \boxed{B}, \boxed{C}, \boxed{D}, \boxed{E}, \boxed{F}, \boxed{G}, \boxed{H}, \boxed{I}, \boxed{J}, \boxed{K}, \boxed{L}$$

However, since person A must be seated with person B , we can regard them as one element. We accomplish this by "bundling" A and B together or alternatively gluing A and B side-by-side to make sure they are together. The picture now looks like this:

$$\boxed{AB}, \boxed{C}, \boxed{D}, \boxed{E}, \boxed{F}, \boxed{G}, \boxed{H}, \boxed{I}, \boxed{J}, \boxed{K}, \boxed{L}$$

Every arrangement of the 11 blocks above will insure that persons A and B are together. Thus, we proceed in steps:

1) Arrange the 11 blocks above. There are $11!$ ways to do so.

2) Within every arrangement of the 11 blocks, we can switch persons A and B . That is, we can either have AB or BA . There are 2 choices here.

By The Basic Principle of Counting, we have a total of $11! \times 2 = 79,833,600$ different arrangements where persons A and B are seated together.

2.3: [Permutations](#) is shared under a [CC BY](#) license and was authored, remixed, and/or curated by LibreTexts.

- **7.3: Permutations** by [Rupinder Sekhon and Roberta Bloom](#) is licensed [CC BY 4.0](#). Original source: <https://www.deanza.edu/faculty/bloomroberta/math11/afm3files.html.html>.

2.3.1: The Multinomial Part 1

With the topic of permutations under our belt, we will now consider the following question.

? Critique This Solution 2.3.1.1

Suppose we have two identical copies of Homer's *Odyssey* and one copy of Plato's *Republic*. How many ways are there to arrange these three books on a bookshelf?

Critique This Solution:

We have three objects and by our [theorem](#) from the last section, there are $3! = 3 \times 2 \times 1 = 6$ different ways to arrange 3 objects. Is this correct? Why or why not?

Answer

The above answer is incorrect! If we were given three *distinct* objects, then there are certainly $3!$ ways to arrange 3 distinct objects. However, we do not have three distinct (or unique) objects in our question. Notice that some of the objects are identical to each other.

In order to answer the above question, allow us to enumerate all the possibilities. We will denote each copy of Homer's *Odyssey* by H and we will denote the single copy of Plato's *Republic* by P . Enumerating yields the following:

1. (H, H, P)
2. (H, P, H)
3. (P, H, H)

We see that there are only three possible arrangements.

The point of the above example is to illustrate that as soon as we break one of our assumptions in the last section (namely that we have distinct elements), our entire theory seems to fall apart. The goal of this mini-section is to develop a formula which tells us how many ways we can arrange n objects when some of the objects are identical to each other.

In order to answer the above question, we first make a definition.

Definition: Type

Objects are said to be of the same **type** if the objects are identical/indistinguishable from each other.

For instance, in the opening exercise, the H and H are of the same type since each copy of Homer's *Odyssey* is identical to each other. We now present the answer to our question in the form of a theorem. The proof of the theorem is discussed at the end of this section.

Theorem 2.3.1.2

Given n objects where:

- n_1 objects are of Type 1,
- n_2 objects are of Type 2,
- ...
- n_k objects are of Type k ,

and $n_1 + n_2 + \dots + n_k = n$, the number of different ways to arrange the n objects is given by

$$\frac{n!}{n_1!n_2!\dots n_k!}$$

✓ Example 2.3.1.3

Suppose we have two identical copies of Homer's *Odyssey* and one copy of Plato's *Republic*. How many ways are there to arrange these three books on a bookshelf?

Answer

In this problem, we have two objects of Type 1 (where Type 1 is Homer's *Odyssey*) and one object of Type 2 (where Type 2 is Plato's *Republic*). By our [theorem](#), there are

$$\frac{3!}{2!1!} = 3$$

different arrangements.

✓ Example 2.3.1.4

How many different arrangements are there of the word AAABB?

Answer

In this problem, we have three objects of Type 1 (where Type 1 are the A's) and two objects of Type 2 (where Type 2 are the B's). By our [theorem](#), there are

$$\frac{5!}{3!2!} = 10$$

different arrangements.

Allow us to explore some additional questions related to the example above because it may be helpful in understanding the the proof of the [theorem](#).

✓ Example 2.3.1.5

1. List of all arrangements of the word AAABB.
2. Suppose the objects of Type 1 were distinct. That is, assume all of the A's are distinct (we can achieve this by labeling or indexing the A's). How many different arrangements are there now?

Answer

Here are all of the arrangements of AAABB:

(A,A,A,B,B)

(A,A,B,A,B)

(A,B,A,A,B)

(B,A,A,A,B)

(B,A,A,B,A)

(B,A,B,A,A)

(B,B,A,A,A)

(A,A,B,B,A)

(A,B,B,A,A)

(A,B,A,B,A)

If the A's were all distinct, then we have to arrange the following: A_1, A_2, A_3, B, B . For convenience, we will label them as (A1, A2, A3, B, B). We can do this neatly via a grid:

	A1, A2, A3	A1, A3, A2	A2, A1, A3	A2, A3, A1	A3, A1, A2	A3, A2, A1
(A,A,A,B,B)	(A1,A2,A3,B,B)	(A1,A3,A2,B,B)	(A2,A1,A3,B,B)	(A2,A3,A1,B,B)	(A3,A1,A2,B,B)	(A3,A2,A1,B,B)
(A,A,B,A,B)	(A1,A2,B,A3,B)	(A1,A3,B,A2,B)	(A2,A1,B,A3,B)	(A2,A3,B,A1,B)	(A3,A1,B,A2,B)	(A3,A2,B,A1,B)
(A,B,A,A,B)	(A1,B,A2,A3,B)	(A1,B,A3,A2,B)	(A2,B,A1,A3,B)	(A2,B,A3,A1,B)	(A3,B,A1,A2,B)	(A3,B,A2,A1,B)
(B,A,A,A,B)	(B,A1,A2,A3,B)	(B,A1,A3,A2,B)	(B,A2,A1,A3,B)	(B,A2,A3,A1,B)	(B,A3,A1,A2,B)	(B,A3,A2,A1,B)
(B,A,A,B,A)	(B,A1,A2,B,A3)	(B,A1,A3,B,A2)	(B,A2,A1,B,A3)	(B,A2,A3,B,A1)	(B,A3,A1,B,A2)	(B,A3,A2,B,A1)
(B,A,B,A,A)	(B,A1,B,A2,A3)	(B,A1,B,A3,A2)	(B,A2,B,A1,A3)	(B,A2,B,A3,A1)	(B,A3,B,A1,A2)	(B,A3,B,A2,A1)
(B,B,A,A,A)	(B,B,A1,A2,A3)	(B,B,A1,A3,A2)	(B,B,A2,A1,A3)	(B,B,A2,A3,A1)	(B,B,A3,A1,A2)	(B,B,A3,A2,A1)
(A,A,B,B,A)	(A1,A2,B,B,A3)	(A1,A3,B,B,A2)	(A2,A1,B,B,A3)	(A2,A3,B,B,A1)	(A3,A1,B,B,A2)	(A3,A2,B,B,A1)
(A,B,B,A,A)	(A1,B,B,A2,A3)	(A1,B,B,A3,A2)	(A2,B,B,A1,A3)	(A2,B,B,A3,A1)	(A3,B,B,A1,A2)	(A3,B,B,A2,A1)
(A,B,A,B,A)	(A1,B,A2,B,A3)	(A1,B,A3,B,A2)	(A2,B,A1,B,A3)	(A2,B,A3,B,A1)	(A3,B,A1,B,A2)	(A3,B,A2,B,A1)

We see that there are a total of 60 different arrangements of the word A_1, A_2, A_3, B, B . The 60 stems from the fact that there are initially 10 arrangements of the word when the A's are identical. Once the A's are no longer identical, there are 3! ways to permute them and so there are a total of $10 \times 3! = 60$ different arrangements.

✓ Example 2.3.1.6

In the Olympics, there are 10 competitors. Of the 10 competitors, 4 competitors represent the United States, 3 competitors represent China, 2 competitors represent England and 1 competitor represents Canada. If the results just lists out the countries of the players in the order that they placed, then how many outcomes are possible?

Answer

According to the question, an outcome is a ranking of the countries. That is, an outcome looks like (United States, England, Canada, England, China, China, China, United States, United States, United States). We see that we have four objects of Type 1 (the competitors representing The United States), 3 objects of Type 2 (the competitors representing China), 2 objects of Type 3 (the competitors representing England), and 1 objects of Type 4 (the competitor representing Canada). By our [theorem](#), there are

$$\frac{10!}{4!3!2!1!} = 12,600$$

different rankings.

We end this section with a seemingly unimportant result followed by the proof of the [theorem](#).

Note

Suppose we have n elements where k elements of Type 1 and $n - k$ elements are of Type 2.
How many different ways can we arrange the n elements?

Answer

By our [theorem](#), there are

$$\frac{n!}{k!(n-k)!}$$

different arrangements. We may see this expression again in the next section!

Theorem 2.3.1.7

Given n objects where:

- n_1 objects are of Type 1,
- n_2 objects are of Type 2,
- ...
- n_k objects are of Type k ,

and $n_1 + n_2 + \dots + n_k = n$, the number of different ways to arrange the n objects is given by

$$\frac{n!}{n_1!n_2!\dots n_k!}$$

Proof

Let d denote the number of distinct arrangements of the n objects. We wish to show that $d = \frac{n!}{n_1!n_2!n_3!\dots n_k!}$.

Notice that if the objects of Type 1 were distinct (which can be achieved by labeling or making an index), then there are $d \times n_1!$ different arrangements.

Also, if the objects of Type 2 were distinct (which can be achieved by labeling or making an index), then there are $d \times n_1! \times n_2!$ different arrangements.

Also, if the objects of Type 3 were distinct (which can be achieved by labeling or making an index), then there are $d \times n_1! \times n_2! \times n_3!$ different arrangements.

...

Finally, if the objects of Type k were distinct, (which can be achieved by labeling or making an index), then there are $d \times n_1! \times n_2! \times n_3! \times \dots \times n_k!$ different arrangements.

At this point in time, we have indexed a total of $n_1 + n_2 + n_3 + \dots + n_k$ objects. By assumption, $n_1 + n_2 + n_3 + \dots + n_k = n$ and so at this point in time, all n objects are distinct.

Thus, we have counted the number of different arrangements of these n distinct objects in two different ways. One way is $d \times n_1! \times n_2! \times n_3! \times \dots \times n_k!$ and the other way is $n!$.

Since these expressions count the same quantity, they are equal to each other. That is, $d \times n_1! \times n_2! \times n_3! \times \dots \times n_k! = n!$. Solving for d yields

$$d = \frac{n!}{n_1!n_2!\dots n_k!}$$

which is precisely what we needed to show.

CHAPTER OVERVIEW

Chapter 3: Week 3

[3.1: Combinations](#)

[3.2: The Multinomial](#)

[3.2.1: The Multinomial Part 2](#)

[Chapter 3: Week 3](#) is shared under a [CC BY](#) license and was authored, remixed, and/or curated by LibreTexts.

3.1: Combinations

As mentioned last time, the following flowchart summarizes how each structure is naturally derived from the one before it:

In the last section, we saw The Generalized Principle of Counting leads to a permutation. In this section, we will discuss how permutations leads to the idea of a combination.

In order to discuss the theory of combinations, we first define what a combination is.

Definition: Permutation

Suppose a set has n distinct elements and an experiment consists of selecting k of the elements one at a time without replacement. Let each outcome consist of the k elements in any order. Each such outcome is called a combination of n elements taken k at a time or is called an unordered sample without replacement.

Moreover, if the k elements selected are $s_1, s_2, s_3, \dots, s_k$, then we denote this combination of n elements taken k at a time by the following notation: $\{s_1, s_2, s_3, \dots, s_k\}$.

The total number of combinations of n elements taken k at a time is denoted by $C_{n,k}$ or $\binom{n}{k}$ which is read as " n choose k ".

There is a lot to take in the above definition and so allow us to highlight the important assumptions above.

Assumptions of a Permutation

In order to have a combination, the following assumptions must be met:

1. We have n distinct elements.
2. We are selecting k of the elements.
3. The selection takes place without replacement (meaning there is no repetition of elements).
4. For some reason, we do not wish to note the order in which the elements are selected. That is, there is nothing special/unique about being chosen first, second or third, and so on.

Once all four assumptions are met, we can denote the combination by writing the elements between braces as we would normally do with our set notation.

Here is a classic example of a scenario involving a combination.

Example 1 3.1.1

Suppose that in a group of 26 friends, I am inviting three friends to attend a concert. Explain why the three people we invite can be thought of as a combination of 26 elements taken 3 at a time. Can you write down a few different outcomes?

Solution

In order to explain why our invitee list can be thought of as combination of 26 elements taken 3 at a time, we simply verify that all four assumptions outlined above apply to the given situation.

1. We have 26 distinct friends.
2. We are selecting 3 of the friends.
3. The selection takes place without replacement since once we choose a friend to attend the concert with us, we cannot choose that friend again. Hence, there is no repetition.
4. The order of selection is irrelevant because if we were to first choose Alice then Bob then Charlie, this is the same as first choosing Charlie, then Bob then Alice. In each case, I am attending the concert with the same three people and so there is nothing special/unique about being chosen first, second or third.

Hence an outcome can be viewed as a combination of 26 elements taken 3 at a time.

Now suppose my friends are labeled $A, B, C, D, \dots, Z$. Here is an example of a combination of 26 elements taken 3 at a time: $\{C, A, R\}$. In this outcome, friends C, A and R were invited to the concert.

Here is another (different!) outcome: $\{K, A, R\}$. This outcome is different because here, friend K is attending whereas in the previous outcome, K was not attending.

Finally, although this was not asked in the problem, note that $\{B, T, B\}$ is not a possible outcome since in a combination, there is no repetition.

📌 Intuition of a Combination

Informally speaking, a combination of n elements taken k at a time is just simply a subset of k elements without replacement. For instance, a combination of n elements taken 2 at a time is just simply a subset of two elements, $\{x, y\}$. Meanwhile, a permutation of n elements taken 3 at a time is just simply a subset of three elements, $\{x, y, z\}$, so on and so forth.

With this idea in mind, allow us to try and count the number of combinations of n elements taken k at a time. We first start off with a concrete example and ask some questions concerning the problem.

✓ Example 2 3.1.2

Find the total number of combinations of 4 elements taken 2 at a time.

Critique This Solution

In order to answer this question, we first must understand what the problem is asking us. Based off our intuition of a combination, this problem is asking us to find the total number of subsets of size 2 without replacement. This means we must find the number of ways to fill in the following structure: $(\{_, _ \})$.

We can think of filling in the above spaces as a 3-stage experiment. We proceed in steps:

1. Choose an element for the first placeholder: there are 4 choices.
2. Choose an element for the second placeholder 2: there are 3 choices.
3. Choose an element for the third placeholder: there are 2 choices.

By The Generalized Principle of Counting, there are $4 \times 3 \times 2 = 12$ different combinations.

Do you think the above solution is correct? Why or why not?

Answer

Unfortunately, the above solution is incorrect! It is incorrect to use The Generalized Principle of Counting. Recall that The Generalized Principle of Counting requires us to have distinct steps! In the proposed solution, there is no difference between step 1, 2 or 3 since we are filling out a combination. In a combination, we said the order in which we select the elements does not matter and so we should not select the elements in separate steps. Thus the proposed solutions are incorrect. In order to correctly answer this problem, we need a different technique.

With the above example in mind, allow us to consider the more general problem:

? Example 3 3.1.3

Find the total number of combinations of 4 elements taken 2 at a time.

Answer

For this question, we seem to be stuck. Whenever we are stuck in a counting question, there is one technique which will always work (although it can be time-consuming and difficult!). Specifically, the technique we will use to answer this question is enumeration. This means we are going to list out all of the possibilities.

As mentioned above, this problem is asking us to find the total number of subsets of size 2 without replacement. This means we must find the number of ways to fill in the following structure: $(\{_, _ \})$ Listing out all possibilities yields the following:

$\{a, b\}$
$\{a, c\}$
$\{a, d\}$
$\{b, c\}$

$\{b, d\}$
 $\{c, d\}$

We see that there are a total of 6 different combinations of 4 elements taken 2 at a time. In terms of notation, we have found that $\binom{4}{2} = 6$.

Unfortunately, the above process does not tell us anything on how to generalize the problem to the case of n elements taken k at a time. And so we must make another insight concerning the problem. What is this insight? Allow us to consider the following related problem.

? Example 4.3.1.4

Enumerate all *permutations* of 4 elements taken 2 at a time.

Answer

For this question, it is clear that there should be $P_{4,2} = 4 \times 3 \times 2 = 12$ such permutations. What are these permutations? Here they are:

 $(a, b), (b, a)$
 $(a, c), (c, a)$
 $(a, d), (d, a)$
 $(b, c), (c, b)$
 $(b, d), (d, b)$
 $(c, d), (d, c)$

How does the above solution help us? Let us create a chart to see if we notice anything:

Combinations of 4 elements taken 2 at a time	Permutations of 4 elements taken 2 at a time
$\{a, b\}$	$(a, b), (b, a)$
$\{a, c\}$	$(a, c), (c, a)$
$\{a, d\}$	$(a, d), (d, a)$
$\{b, c\}$	$(b, c), (c, b)$
$\{b, d\}$	$(b, d), (d, b)$

$\{c, d\}$
 $(c, d), (d, c)$

Looking at the above chart, is there anything that we notice? That is, is there any formula we can develop between the number of combinations of 4 elements taken 2 at a time and permutations of 4 elements taken 2 at a time?

Answer

- 1) The first thing we notice is that there are 6 entries on the left hand side and 12 entries on the right hand side.
 - 2) Where does the 6 come from? The 6 stems from the fact that we computed $\binom{4}{2} = 6$. Similarly, where does the 12 come from? The 12 stems from the fact that $P_{4,2} = 12$.
 - 3) There are twice as many permutations as combinations and so $\binom{4}{2} \times 2 = P_{4,2}$. But where did the 2 come from?
 - 4) To answer where the 2 came from, we note the once difference between a combination and a permutation. As seen from our list of assumptions, in a combination the order is not noted while for a permutation, the order is indeed noted. *Thus, given any combination, if we were to impose an ordering, then we would obtain a permutation!* This is precisely what is happening in our chart. Fix any row and start with a combination. If we were to impose an ordering on the two elements, then we would end up with the corresponding permutation. The question now becomes, how many ways can we order two elements? By our theorem from Section 2.3, there are $2!$ ways to arrange two distinct elements.
 - 5) Putting everything together, it seems like $\binom{4}{2} \times 2! = P_{4,2}$ where the $2!$ stems from the fact that there are $2!$ ways to arrange 2 distinct elements.
- If we had to guess, the general formula which links permutations and combinations is given by $C_{n,k} \times k! = P_{n,k}$ and hence

$$C_{n,k} = \frac{P_{n,k}}{k!}$$

Note

Based off the above discussion, we expect $C_{n,k} \times k! = P_{n,k}$. If you are not convinced, then we can do the same procedure with different numbers. For instance, let us check if $\binom{5}{3} \times 3! = P_{5,3}$. Clearly $P_{5,3} = 60$. To find $\binom{5}{3}$ try enumerating all the possibilities. If you get stuck, then run the following script:

```
combn(5, 3, FUN = NULL, simplify = FALSE)
```


If done correctly, we will see that $\binom{5}{3} \times 3!$ does indeed equal $P_{5,3}$.

We can now finally answer the following question: Find the number of combinations of n elements taken k at a time. The answer is presented in the form of the following theorem.

Theorem: Total Number of Combinations 3.1.5

$$C_{n,k} = \frac{P_{n,k}}{k!} = \frac{n!}{k!(n-k)!}$$

Proof: We consider how a permutation is constructed. Specifically, a permutation is constructed in two steps.

- 1) A subset of k elements are chosen. The number of ways to generate a subset of k elements without repetition is given by $C_{n,k}$.
- 2) We then order the k elements. There are $k!$ ways to do so.

The above two-stage process yields all possible permutations, $P_{n,k}$. Hence, by The Basic Principle of Counting $C_{n,k} \times k! = P_{n,k}$ and so

$$C_{n,k} = \frac{P_{n,k}}{k!}$$

Note that Substituting in $\frac{n!}{(n-k)!}$ for $P_{n,k}$ in $\frac{P_{n,k}}{k!}$ yields $\frac{n!}{k!(n-k)!}$.

We now add on a question to the first example we discussed at the beginning of this section.

✓ Example 5 3.1.5

Suppose that in a group of 26 friends, I am inviting three friends to attend a concert. Find the number of different friend groups that I can invite.

Answer

As we discussed in the first example, each group of three friends can be regarded as a subset of three people. Hence, the question is asking us to find the number of subsets of size three. By the above theorem, the number of different combinations of friend groups is given by

$$C_{26,3} = \frac{P_{26,3}}{3!} = 15,600$$

✓ Example 6 3.1.6

From a group of 28 people, two people are chosen to be secretaries. How many different choices are possible?

Answer

We first ask ourselves what counting structure do we have here? Observe the following:

- 1) Do we have n distinct elements? Yes, we have 28 distinct people.
- 2) In this experiment are we selecting k elements? Yes, we are selecting 2 people.
- 3) Does the selection happen without replacement? Yes, the selection happens without replacement (otherwise it's possible we pick the the same person both times and so we only have 1 secretary as opposed to 2).
- 4) Is the order of selection noted? In order to answer this, we ask ourselves is there anything special about being chosen first versus being chosen second? The answer is no! Both chosen people end up with the same task/role of being a secretary. This means we may think of an outcome as a subset of two people, $\{\underbrace{_, _}_{\text{sec.}}\}$.

by $C_{28,2} = \binom{28}{2} = 378$.

Let us now reconsider the problem posed in our last section:

✓ Example 7: Critique This Solution! 3.1.7

Suppose that from a group of 30 people, a president, vice president, and two secretaries are to be chosen. How many different outcomes are possible?

Answer

For this question, we can consider an outcome to look like the following: $(\underbrace{_, _}_{\text{pres.}}, \underbrace{_, _}_{\text{vp.}}, \underbrace{_, _}_{\text{sec.}})$. That is, we can think of an

outcome as being fulfilled in three steps:

- 1) We first select a president. There are 30 choices.
- 2) We then select a vice president. There are 29 choices.
- 3) We then select a subset of two people to be secretaries. There are $\binom{28}{2}$ choices.

By The Generalized Principle of Counting, there are $30 \times 29 \times \binom{28}{2} = 328,860$ different outcomes.

✓ Example 8 3.1.8

From a group of 10 women and 12 men, a committee of 2 women and 3 men must be formed. How many different committees are possible?

Answer

For this question, we can consider an outcome to look like the following: $\{\underbrace{_, _, _}_{\text{women}}, \underbrace{_, _, _}_{\text{men}}\}$. That is, we can think of an outcome as a subset of 5 people where 2 people are women and the other 3 are men. Thus, an outcome is fulfilled in two steps:

- 1) We first choose a subset of two women. There are $\binom{10}{2} = 45$ ways to do this.
- 2) We now select a subset of three men. There are $\binom{12}{3} = 220$ ways to do this.

By The Generalized Principle of Counting, there are $\binom{10}{2} \times \binom{12}{3} = 9,900$ different committees consisting of 2 women and 3 men.

✓ Example 9 3.1.9

From a regular deck of 52 cards, 5 cards are drawn at random with no emphasis on the order of the cards. How many outcomes are there where we obtain 2 Kings, 2 Queens, and 1 Jack?

Answer

For this question, we can consider an outcome to look like the following: $\{\underbrace{_, _}_{\text{Kings}}, \underbrace{_, _}_{\text{Queens}}, \underbrace{_}_{\text{Jacks}}\}$. That is, we can think of an outcome as a subset of 5 cards where 2 cards are Kings, 2 cards are Queens, and the other card is a Jack. Thus, an outcome is fulfilled in three:

- 1) We first choose a subset of two Kings. There are $\binom{4}{2} = 6$ ways to do this.
- 2) We now select a subset of two Queens. There are $\binom{4}{2} = 6$ ways to do this.
- 3) We select one Jack. There are $\binom{4}{1} = 4$ ways to do this.

By The Generalized Principle of Counting, there are $\binom{4}{2} \times \binom{4}{2} \times \text{binom}41 = 4 = 144$ different subsets of consisting of 2 Kings, 2 Queens, and 1 Jack.

✓ Example 10 3.1.9

From a regular deck of 52 cards, 5 cards are drawn at random with no emphasis on the order of the cards. How many outcomes are there where we obtain three 7's and two Aces?

Answer

For this question, we can consider an outcome to look like the following: $\{\underbrace{_, _, _}_{\text{Sevens}}, \underbrace{_, _}_{\text{Aces}}\}$. That is, we can think of an outcome as a subset of 5 cards where 3 of the cards have the denomination of a 7 and the other two cards have the denomination of an Ace. We proceed in steps.

- 1) We first choose three cards that are labeled with a seven. There are $\binom{4}{3} = 4$ ways to do this.
- 2) We now choose two cards that are labeled with an ace. There are $\binom{4}{2} = 6$ ways to do this.

By The Generalized Principle of Counting, there are $\binom{4}{3} \times \binom{4}{2} = 24$ different subsets of consisting of three 7's and two Aces.

✓ Example 10 3.1.10

From a regular deck of 52 cards, 5 cards are drawn at random with no emphasis on the order of the cards. How many outcomes are there where we obtain a full house? (By a *full house* we mean that we obtain three cards of one denomination and two cards of another denomination, different from the first. For instance, an example of a full house would be $7\heartsuit, 7\diamondsuit, 7\clubsuit, A\heartsuit, A\spadesuit$.)

Answer

For this question, we can consider an outcome to look like the following: $\{\underbrace{_, _, _}_{\text{Three of a kind}}, \underbrace{_, _}_{\text{Two of a kind}}\}$. That is, we can think of an outcome as a subset of 5 cards where 3 of the cards have the same denomination and the other two cards have the same denomination. We proceed in steps.

- 1) We first choose a denomination for the three of a kind. There are $\binom{13}{1} = 13$ ways to do this.

- 2) For the selected denomination, we choose three of the four cards. There are $\binom{4}{3} = 4$ ways to do this.
 - 3) We now choose a denomination for the two of a kind. There are $\binom{12}{1} = 12$ ways to do this.
 - 4) For the selected denomination in step 3, we choose two of the four cards. There are $\binom{4}{2} = 6$ ways to do this.
- By The Generalized Principle of Counting, there are $\binom{13}{1} \times \binom{4}{3} \times \binom{12}{1} \times \binom{4}{2} = 3,744$ different full houses.

An important note to make is that for certain problems, there is not one unique set of steps to a solution. For instance, consider the following:

? Exercise 11 - Alternative Solution to Example 10 3.1.11

In the above example, notice that the answer $\binom{13}{2} \binom{2}{1} \binom{4}{3} \binom{4}{2}$ is equivalent to 3,744. Can you determine the steps associated with this alternative solution?

Answer

For this question, we still regard an outcome as the following: $\{ \underbrace{\quad, \quad, \quad}_{\text{Three of a kind}}, \underbrace{\quad, \quad}_{\text{Two of a kind}} \}$. The only thing which is different

is the order of our steps:

- 1) We first choose the two denominations that will be in our full house. There are $\binom{13}{2} = 78$ ways to do this.
 - 2) From these two denominations, we must choose one denomination that will be used to fill in the three of a kind. There are $\binom{2}{1} = 2$ ways to do this.
 - 3) For the selected denomination, we choose three of the four cards. There are $\binom{4}{3} = 4$ ways to do this.
 - 4) For the other denomination, we choose two of the four cards. There are $\binom{4}{2} = 6$ ways to do this.
- By The Generalized Principle of Counting, there are $\binom{13}{2} \binom{2}{1} \binom{4}{3} \binom{4}{2} = 3,744$ different full houses.

We end this section with

Theorem 3.1.1

$$\binom{n}{k} = \binom{n}{n-k}$$

Proof: The formal proof is left as a homework exercise.

Intuition (Informal Justification of the above theorem)

For example, the above theorem is stating that

$$\binom{30}{5} = \binom{30}{25}$$

Why is this true? Let us suppose we have 30 students in a room and we have to select 5. Clearly, $\binom{30}{5}$ represents the number of subsets of size 5. But notice every time we choose a subset of 5 students what are we also doing? We are *not* choosing 25 students. Hence, for every subset of 5 students, we are indirectly creating a subset of 25 students and vice-versa. Each time we choose 25 students, we are indirectly creating a subset of 5 students. Thus

$$\binom{30}{5} = \binom{30}{25}$$

. More generally,

$$\binom{n}{k} = \binom{n}{n-k}$$

since every time we choose k elements, we are creating a subset of $n - k$ elements and vice-versa.

3.2: The Multinomial

With the topic of combinations under our belt, we will now consider the following question.

? Example 1 3.2.1

In a hospital, 14 nurses are to be divided so that 5 nurses are in Unit 1 (Cardiac Unit), 4 nurses are in Unit 2 (Emergency Room), 3 nurses are in Unit 3 (Medical Surgery) and 2 nurses are in Unit 4 (Labor and Delivery). How many different ways are there to divide the 14 nurses?

Answer

We can think of an outcome as having the following structure: $(\underbrace{-, -, -, -, -}_{\text{Cardiac.}}, \underbrace{-, -, -, -}_{\text{ER}}, \underbrace{-, -, -}_{\text{MedSurg.}}, \underbrace{-, -}_{\text{Labor}})$. We now proceed in steps:

1. We first choose 5 nurses to be assigned to the Cardiac unit. There are $\binom{14}{5}$ ways to do this.
2. We now select 4 nurses to be assigned to the Emergency Room. There are $\binom{9}{4}$ ways to do this.
3. We then select 3 nurses to be assigned to the Medical Surgery unit. There are $\binom{5}{3}$ ways to do this.
4. We then select 2 nurses to be assigned to the Labor and Delivery unit. There are $\binom{2}{2}$ ways to do this.

By The Generalized Principle of Counting, there are $\binom{14}{5} \times \binom{9}{4} \times \binom{5}{3} \times \binom{2}{2} = 2,522,520$ different divisions.

While this answers the question, notice the following:

$$\binom{14}{5} \times \binom{9}{4} \times \binom{5}{3} \times \binom{2}{2} = \frac{14!}{5!9!} \times \frac{9!}{4!5!} \times \frac{5!}{3!2!} \times \frac{2!}{2!0!} = \frac{14!}{5!4!3!2!}$$

Does the term $\frac{14!}{5!4!3!2!}$ seem familiar?

This resembles the number of ways to arrange 14 elements when 5 objects are of Type 1, 4 objects are of Type 2, 3 objects are of Type 3 and 2 objects are of Type 4.

We will now generalize the above result.

🔗 Theorem 3.2.2

If n distinct elements are to be divided among k distinct groups such that:

- Group 1 has n_1 elements,
- Group 2 has n_2 elements,
- ...
- Group k has n_k elements,

and $n_1 + n_2 + \dots + n_k = n$, then the number of different divisions is given by

$$\frac{n!}{n_1!n_2!\dots n_k!}$$

Proof: We proceed in steps:

1. We first choose n_1 elements to be assigned to Group 1. There are $\binom{n}{n_1}$ ways to do this.
2. We first choose n_2 elements to be assigned to Group 2. There are $\binom{n-n_1}{n_2}$ ways to do this.
3. We first choose n_3 elements to be assigned to Group 3. There are $\binom{n-n_1-n_2}{n_3}$ ways to do this.
- ...
- k. We first choose n_k elements to be assigned to Group k . There are $\binom{n-n_1-n_2-\dots-n_{k-1}}{n_k}$ ways to do this.

By The Generalized Principle of Counting, there are

$$\binom{n}{n_1} \times \binom{n-n_1}{n_2} \times \binom{n-n_1-n_2}{n_3} \times \dots \times \binom{n-n_1-n_2-\dots-n_{k-1}}{n_k} = \frac{n!}{n_1!(n-n_1)!} \times \frac{(n-n_1)!}{n_2!(n-n_1-n_2)!} \times \frac{n_2!(n-n_1-n_2)!}{n_3!(n-n_1-n_2-n_3)!} \times$$

$$\dots \frac{(n-n_1-n_2-\dots-n_{k-1})!}{n_k!(n-n_1-n_2-\dots-n_{k-1}-n_k)!} = \frac{n!}{n_1!n_2!\dots n_k!}$$
divisions.

3.2: The Multinomial is shared under a [CC BY](#) license and was authored, remixed, and/or curated by LibreTexts.

3.2.1: The Multinomial Part 2

3.2.1: The Multinomial Part 2 is shared under a [CC BY](#) license and was authored, remixed, and/or curated by LibreTexts.

CHAPTER OVERVIEW

Chapter 4: Week 4

4.1: Simple Sample Spaces

4.2: Overcounting

Chapter 4: Week 4 is shared under a [CC BY](#) license and was authored, remixed, and/or curated by LibreTexts.

4.1: Simple Sample Spaces

Recall that the main topic of the last few sections was combinatorial analysis. In this section, we relate how counting is related to probability. We first must make the following two definitions.

Definition: Cardinality

If A is a finite set, then we define the cardinality of A , denoted by $|A|$, to be the number of elements in A .

✓ Example 1 4.1.1

Compute the cardinalities of the following sets:

1. $S = \{H, T\}$
2. $S = \{1, 2, 3, 4, 5, 6\}$
3. $S = \{\text{all possible selections of a president, vice president, secretary and treasurer from a group of 30 people}\}$.

Solution

1. $|S| = 2$
2. $|S| = 6$
3. $|S| = P_{30,4}$

Definition: Simple Sample Spaces

A simple sample space refers to a finite sample space where all of the outcomes are equally likely.

✓ Example 2 4.1.1

1. Suppose an experiment consists of flipping a fair coin. Then $S = \{H, T\}$. This sample space is finite and since the coin is fair, both outcomes are equally likely. Hence, under this experiment, S is a simple sample space.
2. Suppose an experiment consists of rolling a regular 6-sided (fair) die. Then $S = \{1, 2, 3, 4, 5, 6\}$. This sample space is finite and since the coin is fair, both outcomes are equally likely. Hence, under this experiment, S is a simple sample space.
3. Suppose an experiment consists of rolling a 6-sided *loaded* die. That is, the die is weighted in such a way that it is more likely to roll a 6 as opposed to a 1. In this example, S is still $\{1, 2, 3, 4, 5, 6\}$. However, under this experiment, S is not a simple sample space since the outcome of rolling a 6 is more likely than rolling a 1.

We will now discuss an important result concerning simple sample spaces which will be our focus for the remainder of this section.

Theorem 4.1.1

If $S = \{s_1, s_2, \dots, s_n\}$ is a simple sample space then

$$P(\{s_1\}) = P(\{s_2\}) = \dots = P(\{s_n\}) = \frac{1}{n}$$

Proof

Since $S = \{s_1\} \cup \{s_2\} \cup \dots \cup \{s_n\}$ then

$$P(S) = P(\{s_1\} \cup \{s_2\} \cup \dots \cup \{s_n\})$$

By assumption, all of the n elements in S are mutually exclusive and so by Axiom 3, we have

$$P(S) = P(\{s_1\}) + P(\{s_2\}) + \dots + P(\{s_n\}) \quad (4.1.1)$$

By the definition of a simple sample space, all of the outcomes are equally likely and so we will denote their probabilities by lowercase p . That is, $P(\{s_1\}) = P(\{s_2\}) = \dots = P(\{s_n\}) = p$ and so Equation 4.1.1 becomes

$$P(S) = \underbrace{p + p + \dots + p}_{n \text{ times}}$$

$$1 = np$$

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

Theorem 4.1.1

If $S = \{s_1, s_2, \dots, s_n\}$ is a simple sample space and A is any event in S such that $|A| = m$, then

$$P(A) = \frac{m}{n} = \frac{|A|}{|S|} \quad (4.1.2)$$

Proof

Since $|A| = m$ then A has the following form $A = \{s_{i_1}, s_{i_2}, \dots, s_{i_m}\}$. Thus, we have

$$P(A) = P(\underbrace{s_{i_1} \cup s_{i_2} \cup \dots \cup s_{i_m}}_{\text{by disjointness}})$$

$$P(A) = \underbrace{P(s_{i_1}) + P(s_{i_2}) + \dots + P(s_{i_m})}_{\text{by the above theorem each probability is } 1/n}$$

$$P(A) = \underbrace{\frac{1}{n} + \frac{1}{n} + \frac{1}{n} + \dots + \frac{1}{n}}_{m \text{ times}}$$

$$P(A) = \frac{m}{n}$$

$$P(A) = \frac{|A|}{|S|}$$

In order to answer questions concerning simple sample spaces, we will use the following framework:

Framework for Answering Problems Regarding Simple Sample Spaces

1. First define the event A .
2. Then define the sample space S by saying "Let S be the set of all outcomes where an outcome is _____" and fill in the blank.
3. Argue that S is indeed a simple sample space. We typically do this by asking ourselves, "is there any reason to suspect that a particular outcome in S is more likely than another outcome in S ?" If the answer is no then we are in a simple sample space. If the answer is yes, then we are not in a simple sample space and we should go back to step 2 and consider a different way to define an outcome.
4. Count $|S|$ by proceeding in steps if necessary.
5. Count $|A|$ by proceeding in steps if necessary.
6. Our answer is given by $P(A) = \frac{|A|}{|S|}$

In the first example, we reconsider the problem which we posed in our very first discussion concerning probabilities.

Example 3 4.1.1

Suppose a fair coin is tossed. What is the probability that a heads is obtained?

Answer

Allow us to apply the framework outlined above.

1. Let $A = \{\text{a heads is obtained}\}$.

2. Let S be the set of all outcomes where an outcome is a result of the coin flip. Hence $S = \{H, T\}$ where H denotes a heads was obtained and T denotes a tail was obtained.
3. Is there any reason to suspect that one outcome in S is more likely than another outcome in S ? Since the coin is fair, there is no reason to believe this. Hence we are in a simple sample space.
4. The cardinality of S is simple enough to compute without having to introduce permutations, combinations, or multinomials. Quite simply, $|S| = 2$.
5. Similarly, $|A| = 1$.
6. Hence, $P(A) = \frac{|A|}{|S|} = \frac{1}{2} = 0.5$.

✓ Example 4 4.1.1

From a group of 10 women and 12 men, 5 people are to be selected at random to be on a committee. What is the probability that we obtain 2 women and 3 men?

Answer

Allow us to apply the framework outlined above.

1. Let $A = \{\text{we obtain 2 women and 3 men}\}$.
2. Let S be the set of all outcomes where an outcome is a subset of 5 people.
3. Is there any reason to suspect that one outcome in S is more likely than another outcome in S ? That is, is there any reason to suspect that a certain subset of 5 specific people is more likely than another subset of 5 specific people? Since the selection happens at random, there is no reason to believe this. Hence we are in a simple sample space.
4. The cardinality of S is $|S| = \binom{22}{5}$.
5. To find the cardinality of A , we recall the following example from Section 3.1. Hence, $|A| = \binom{10}{2} \binom{12}{3}$.
6. Hence, $P(A) = \frac{\binom{10}{2} \binom{12}{3}}{\binom{22}{5}} \approx 0.3759$.

In the above example, it was natural for an outcome to be a subset of 5 people since there was nothing special about the order in which we selected the elements. That being said, let us consider the following question: if for some reason, we did care about the order of selection, would the probability be any different? That is, if we regard an outcome as an ordered 5-tuple, then would the probability change? Let us take a closer look:

✓ Example 5 4.1.1

From a group of 10 women and 12 men, 5 people are to be selected at random to be on a committee. What is the probability that we obtain 2 women and 3 men?

Answer

Allow us to apply the framework outlined above.

1. Let $A = \{\text{we obtain 2 women and 3 men}\}$.
2. Let S be the set of all outcomes where an outcome is an ordered 5-tuple.
3. Is there any reason to suspect that one outcome in S is more likely than another outcome in S ? That is, is there any reason to suspect that a certain ordering of 5 specific people is more likely than another ordering of 5 specific people? Since the selection happens at random, there is no reason to believe this. Hence we are in a simple sample space.
4. The cardinality of S is $|S| = P_{22,5}$.
5. To find the cardinality of A , we proceed in steps.
 1. We first choose a subset of two women. There are $\binom{10}{2} = 45$ ways to do this.
 2. We now select a subset of three men. There are $\binom{12}{3} = 220$ ways to do this.
 3. Since an outcome is ordered, we now must order these 5 distinct people. There are $5!$ ways to do so.

6. So $|A| = \binom{10}{2} \times \binom{12}{3} \times 5!$ Hence, $P(A) = \frac{\binom{10}{2} \times \binom{12}{3} \times 5!}{P_{22,5}} \approx 0.3759$ which is the same answer as above.

Notice that

$$\frac{\binom{10}{2}\binom{12}{3}}{\binom{22}{5}} = \frac{\binom{10}{2} \times \binom{12}{3} \times 5!}{P_{22,5}}$$

and so it turns out that it does not matter if we allow the five people to have an ordering - the probability will be the same. More generally, this is always the case for any experiment that involves a random selection of k elements from n . We summarize this in the following remark.

Remark

Whenever an experiment consists of randomly selecting k elements from n elements, we have some flexibility in defining the outcomes of S . We may either

1. let an outcome be a subset of k elements, or
2. let an outcome be an ordered k -tuple.

The choice is completely up to us. Either way, we will obtain the same probability.

The above remark essentially tells us that it is okay to imagine that the experiment has some sort of additional structure where we care about the ordering of the elements. Again, this only applies if the experiment requires us to randomly selecting k elements from n distinct elements. Moreover, notice that the above remark is general in the sense that it does not matter if the n elements are distinct. In such a case where the n elements are not distinct, we can make them distinct by indexing them as previously discussed. In the above example, some may argue that the subset route was simpler. However, this is not always the case. Consider the following problem which we first tackle via permutations.

✓ Example 6 4.1.1

From a group of 8 married couples, 4 people are chosen at random. Find the probability that the five people who are chosen are not related, meaning that the five people do not belong to the same household.

Answer

Allow us to apply our usual framework.

1. Let $A = \{\text{the five chosen people are unrelated}\}$.
2. Let S be the set of all outcomes where an outcome is an ordered 4-tuple.
3. Is there any reason to suspect that one outcome in S is more likely than another outcome in S ? That is, is there any reason to suspect that a certain ordering of 4 specific people is more likely than another ordering of 4 specific people? Since the selection happens at random, there is no reason to believe this. Hence we are in a simple sample space.
4. The cardinality of S is $|S| = P_{16,4}$.
5. To find the cardinality of A , we proceed in steps.

1. We first choose any person. There are 16 ways to do this.
2. We now must choose a person who is not related to the person above. There are 14 ways to do this.
3. We now must choose a person who is not related to any previous selected person. There are 12 ways to do this.
4. We now must choose a person who is not related to any previous selected person. There are 10 ways to do this.

$$6. \text{ So } |A| = 16 \times 14 \times 12 \times 10 \text{ Hence, } P(A) = \frac{16 \times 14 \times 12 \times 10}{P_{16,4}} \approx 0.6154.$$

We now consider the same problem where the outcomes are subsets of four people.

✓ Example 7 4.1.1

From a group of 8 married couples, 4 people are chosen at random. Find the probability that the five people who are chosen are not related, meaning that the five people do not belong to the same household.

Answer

Allow us to apply our usual framework.

1. Let $A = \{\text{the five chosen people are unrelated}\}$.
2. Let S be the set of all outcomes where an outcome is a subset of 4 people.
3. Is there any reason to suspect that one outcome in S is more likely than another outcome in S ? That is, is there any reason to suspect that a certain subset of 4 specific people is more likely than another subset of 4 specific people? Since the selection happens at random, there is no reason to believe this. Hence we are in a simple sample space.
4. The cardinality of S is $|S| = \binom{16}{4}$.
5. To find the cardinality of A , we proceed in steps. (Remember, we need a set of steps that will generate all subsets of 4 people where the 4 people are not related to each other. This can be difficult and requires some thought!)

1. We first choose four couples. There are $\binom{8}{4}$ ways to do this.
2. Within each of the four couples, we will select one person. This will insure that we end up with four people who are not related to each other. There are $\binom{2}{1} \times \binom{2}{1} \times \binom{2}{1} \times \binom{2}{1}$ ways to do this.

$$6. \text{ So } |A| = \binom{8}{4} \times \binom{2}{1} \times \binom{2}{1} \times \binom{2}{1} \times \binom{2}{1} . \text{ Hence, } P(A) = \frac{\binom{8}{4} \times \binom{2}{1} \times \binom{2}{1} \times \binom{2}{1} \times \binom{2}{1}}{\binom{16}{4}} \approx 0.6154.$$

Reflecting back on both solutions, it appears that the permutation route was more straightforward. In general, you should choose the method best-suited to the problem.

✓ Example 8 - Critique This Solution! 4.1.1

If 3 marbles are randomly withdrawn from a bowl consisting of 997 green marbles and 3 red marbles, then what is the probability that one marble is green and the other two are red?

Critique this solution:

Allow us to apply our usual framework.

1. Let $A = \{\text{we obtain 1 green marbles and 2 red marbles}\}$.
2. Let S denote the set of all outcomes where an outcome is the number of green marbles we get after selecting 3 marbles. Then, $S = \{0 \text{ greens}, 1 \text{ green}, 2 \text{ green}, 3 \text{ greens}\}$.
3. Since the selection happens at random, there is no reason to believe that one outcome is more likely than another outcome.
4. $|S| = 4$.
5. $|A| = 1$.
6. Hence $P(A) = \frac{|A|}{|S|} = \frac{1}{4}$.

Is the above solution correct? Why or why not?

Answer

The above answer is incorrect! Just because the selection happens at random, it does not mean that S is a simple sample space. After some reflection it should be clear that based off the composition of the bowl, it is far more likely that we get 3 greens as opposed to 0 greens and so S is not a simple sample space. As we mentioned in our framework, whenever S is not a simple sample space, we should redefine S . Here is how we can do this:

- 1) Let $A = \{\text{we obtain 1 green marbles and 2 red marbles}\}$.
- 2) Notice in this experiment, we are selecting 3 elements at random from 1000 elements. By our remark, we can let an outcome be a subset of 3 elements or we may let an outcome be an ordered 3-tuple. For concreteness, we will let an outcome be a subset of 3 elements*. For example, here are a few possible outcomes: $\{G_{17}, G_{409}, R_2\}$ or $\{G_1, G_{241}, G_{65}\}$ or

$\{R_1, R_2, R_3\}$. Notice it is implied that the elements are indexed or distinct since a combination requires all elements to be distinct.

3) Since the selection happens at random, there is no reason to believe that one subset of three specific marbles is more likely than another subset of three specific marbles. Thus S is a simple sample space.

4) $|S| = \binom{1000}{3}$.

5) To find $|A|$, we proceed in steps. We need to make sure our outcome looks like this: $\{\underbrace{\quad}_{\text{green}}, \underbrace{\quad}_{\text{red}}\}$. We first select one green. There are $\binom{997}{1}$ ways to do this. We then must

select two reds. There are $\binom{3}{2}$ ways to do this. Hence, $|A| = \binom{997}{1} \times \binom{3}{2}$.

6) Putting everything together, we obtain that $P(A) = \frac{|A|}{|S|} = \frac{\binom{997}{1} \times \binom{3}{2}}{\binom{1000}{3}} \approx 0.000018$.

✓ Example 9 - Critique This Solution! 4.1.1

Suppose two fair dice are rolled. What is the probability that we obtain a sum of 7? (You may assume that the dice are regular 6-sided dice).

Critique this solution:

Allow us to apply our usual framework.

1. Let $A = \{\text{we obtain a sum of 7}\}$.
2. Let S denote the set of all outcomes where an outcome is a possible sum that results from rolling a pair of fair dice. Then, $S = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$
3. Since the dice are fair, there is no reason to believe that one outcome is more likely than another outcome.
4. $|S| = 11$.
5. $|A| = 1$.
6. Hence $P(A) = \frac{|A|}{|S|} = \frac{1}{11}$.

Is the above solution correct? Why or why not?

Answer

This is incorrect because S is not a simple sample space. For if it were, then all of the outcomes would be equally likely. This means the chances of getting a sum of 2 would be the same as getting a sum of 7 but clearly this is not the case! There is only one configuration of the dice that would yield a sum of 2. However, there are multiple configurations of the dice that yields a sum of 7. If you are not convinced, then ask yourself if the following is a fair bet. If you roll a sum of 2 then I will give you \$100 dollars but if I roll a sum of 7 then you have to give me a sum of \$100. Clearly, this is not a fair bet since there is only one way you can possibly win but I have multiple ways to win!

Since S is not a simple sample space, our framework tells us that we should re-define S in order to make it a simple sample space. How can we do this? Looking back at the problem, we notice that we are not selecting k elements from n and so the same trick that worked in the above problem no longer applies. In situations like these, our construction of S should be based on what is actually fair in the experiment. For this experiment, the dice themselves are fair and so an outcome should be a specification of the number obtained on each die. That is, we can think of an outcome as an ordered pair $(\underbrace{\quad}_{\text{Die 1}}, \underbrace{\quad}_{\text{Die 2}})$.

1. Let $A = \{\text{we obtain a sum of 7}\}$.
2. Let S denote the set of all outcomes where an outcome is an ordered pair. The first entry in the ordered pair denotes the outcome of rolling Die 1 and the second entry in the ordered pair denotes the outcome of rolling Die 2.
3. S is a simple sample space because there is no reason to believe that a certain ordered pair is more likely than another ordered pair since the dice are fair. Visually we can list out the outcomes in S via the following chart:

	Die 2	1	2	3	4	5	6
Die 1							
1		(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)

	Die 2	1	2	3	4	5	6
2		(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
3		(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
4		(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
5		(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
6		(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

4. $|S| = 6^2$.
5. $|A| = 6$.
6. $P(A) = \frac{|A|}{|S|} = \frac{6}{36} = \frac{1}{6}$.

The following two problems will make use of the theorems we proved in Section 1.3. Before we get into the next problem, allow us to make the following informal comment.

Comment

In Section 2.1, we stated that following two remarks:

- 1) The word "or" corresponds to the set theoretic operation of union and the word "and" corresponds to the set theoretic operation of intersection.
 - 2) In rephrasing an event, we should always ask ourselves "can we describe this event by using the words *or/and*?"
- Sometimes when rephrasing the event, our rephrasing will use the word "or" many times. As such, we will be left with many unions when we describe the event in terms of sets. If the events in the union are disjoint, then we should be able to proceed directly. However, in the case that we have many unions and the events are not disjoint, the computation can be extremely complicated. For such a situation, it is better to proceed indirectly by studying the complement.

Our next problem illustrates the above comment.

✓ Example 10: The Birthday Problem 4.1.1

If $k \leq 365$ people are in a room, then what is the probability that at least two people will have the same birthday?

You may make the following assumptions:

1. The same birthday means the same month and day, not year.
2. No twins, triplets, etc. are present.
3. Each of the 365 days are equally likely to be someone's birthday.
4. The year is not a leap year.
5. If someone's birthday is 2/29 then we will consider their birthday to be 3/01.

Solution

1. Let $A = \{\text{at least two people share the same birthday}\}$. This event can be rephrased as $A = \{2 \text{ people share the same birthday}\} \cup \{3 \text{ people share the same birthday}\} \cup \dots \cup \{k \text{ people share the same birthday}\}$

disjoint! It is possible to have a room of k people where 2 people share the same birthday and then a group of 3 people also share the same birthday. Looking back at A , we see that A is the union of many non-disjoint events and so proceeding directly is ill advised. Rather, we will study its complement. Notice that

$$\begin{aligned}
 A^c &= \{\text{it is not the case that at least two people share the same birthday}\} \\
 &= \{\text{no one shares the same birthday}\} \\
 &= \{\text{everyone has different birthdays}\} \\
 &= \{\text{Person 1 and Person 2 and Person 3 and } \dots \text{ and Person } k \text{ all have different birthdays}\}
 \end{aligned}$$

2. Let S be the set of all outcomes where an outcome is a list of the birthdays for each person. That is, an outcome can be thought of as the following ordered k -tuple: $(\underbrace{\quad}_{\text{Person 1's birthday}}, \underbrace{\quad}_{\text{Person 2's birthday}}, \dots, \underbrace{\quad}_{\text{Person } k\text{'s birthday}})$.

3. Based off the assumptions that we have outlined above, S is a simple sample space since there is no reason to believe that one list of birthdays is more likely than another list of birthdays.

4. $|S| = 365^k$.

5. To find $|A^c|$, we will proceed in steps.

1. Select a birthday for Person 1. There are 365 choices.
2. Select a different birthday for Person 2. There are 364 choices.
3. Select a different birthday for Person 3. There are 363 choices.

....

k. Select a different birthday for Person k . There are $365 - (k - 1) = 365 - k + 1$ choices.

By The Generalized Principle of Counting, $|A^c| = 365 \times 364 \times \dots \times (365 - k + 1) = P_{365, k}$.

6. Hence $P(A^c) = \frac{|A^c|}{|S|} = \frac{P_{365, k}}{365^k}$ and so

$$P(A) = 1 - P(A^c) = 1 - \frac{P_{365, k}}{365^k}$$

Let us plot the $P(A)$ as a function of k .

```
k <- seq(1,365)
plot(k, 1 - choose(365, 365 - k) * factorial(k) / 365 ^ k,
main="The Birthday Problem",
ylab="Probability")
```

run

restart

restart & run all

It seems that the probability approaches 1 rather quickly! Zooming in, we obtain the following graph:

```
k <- seq(1,60)
plot(k, 1 - choose(365, 365 - k) * factorial(k) / 365 ^ k,
main="The Birthday Problem",
ylab="Probability")
```

run

restart

restart & run all

Let us summarize some of the selected probabilities in the form of a chart.

The number of people in the room, k	The probability that at least two people share the same birthday, $P(A)$
1	0
5	0.027
10	0.117
15	0.253
20	0.411
25	0.569
30	0.706
40	0.891
50	0.970

The number of people in the room, k	The probability that at least two people share the same birthday, $P(A)$
60	0.994

✓ Example 11 4.1.1

Ten marbles are randomly withdrawn without replacement from a bowl that contains 10 red, 15 blue, and 20 green marbles. Find the probability that either exactly 4 red marbles or exactly 4 blue marbles are withdrawn.

Answer

As usual, we apply our counting framework while keeping a keen eye out for ways to simplify the event of interest.

1) Let $A = \{\text{we obtain exactly 4 red marbles or 4 blue marbles}\}$. The word *or* indicates that we have a union of two events. As such, let $A_R = \{\text{we obtain exactly 4 red marbles}\}$ and let $A_B = \{\text{we obtain exactly 4 blue marbles}\}$. Then $A = A_R \cup A_B$. Notice that A_R and A_B are not disjoint since it is possible to select ten marbles and obtain 4 reds and 4 blues. However, we only have one union as opposed to many of them as seen in the previous example and so we will deal with this directly. Doing so yields the following:

$$\begin{aligned} P(A) &= P(A_R \cup A_B) = P(A_R) + P(A_B) - P(A_R \cap A_B) \\ &= \frac{|A_R|}{|S|} + \frac{|A_B|}{|S|} - \frac{|A_R \cap A_B|}{|S|} \quad \text{provided we are in a simple sample space} \end{aligned}$$

2) Since this experiment involves a random selection of k elements from n , we have some flexibility in defining S . We may either let an outcome be a subset of ten marbles or we may let an outcome be an ordered 10-tuple. For convenience, we will let an outcome be a subset of 10 marbles.

3) S is a simple sample space since there is no reason to believe a certain subset of 10 specific marbles is more likely than another subset of 10 specific marbles.

4) $|S| = \binom{45}{10}$.

5) $|A_R| = \binom{10}{4} \binom{35}{6}$, $|A_B| = \binom{15}{4} \binom{30}{6}$, and $|A_R \cap A_B| = \binom{10}{4} \binom{15}{4} \binom{20}{2}$

Putting everything together yields:

$$P(A) = \frac{\binom{10}{4} \binom{35}{6}}{\binom{45}{10}} + \frac{\binom{15}{4} \binom{30}{6}}{\binom{45}{10}} - \frac{\binom{10}{4} \binom{15}{4} \binom{20}{2}}{\binom{45}{10}} \approx 0.3438$$

4.1: Simple Sample Spaces is shared under a CC BY license and was authored, remixed, and/or curated by LibreTexts.

4.2: Overcounting

In this section, we will explore a common fallacy that occurs while counting. Allow us to consider (and critique!) the following solution:

✓ Critique This Solution 4.2.1

Poker dice is played by simultaneously rolling 5 fair six-sided dice. Find the probability that we obtain one pair.

Note: By *one pair*, we mean that the dice have denominations a, a, b, c, d where a, b, c and d are all distinct. For instance, the following are examples of one pair: $(2, 1, 2, 3, 5)$ and $(6, 2, 1, 5, 5)$. Meanwhile, the outcomes $(2, 1, 2, 3, 2)$ and $(2, 2, 1, 1, 1)$ do not meet the conditions of being one pair.

Solution

As usual, we apply our counting framework. Let $A = \{\text{we obtain one pair}\}$ and let S be the set of all outcomes where an outcome is an ordered 5-tuple. That is, an outcome looks like $(_, _, _, _, _)$. Notice there is no reason to believe that a certain ordered 5-tuple is more likely than another ordered 5-tuple and so S is indeed a simple sample space. Clearly, $|S| = 6^5$. All that remains is to compute $|A|$. To do this, we proceed in steps:

1. We first choose a value of a . That is, we must select the denomination to be repeated. There are $\binom{6}{1}$ ways to do so.
2. We now must select a value of b that is different from a . There are $\binom{5}{1}$ ways to do so.
3. We now must select a value of c that is different from a and b . There are $\binom{4}{1}$ ways to do so.
4. We now must select a value of d that is different from a, b and c . There are $\binom{3}{1}$ ways to do so.
5. Finally, we must impose an ordering on a, a, b, c, d . There are $\frac{5!}{2! 1! 1! 1!}$ ways to do so.

$$\text{Hence } |A| = \binom{6}{1} \times \binom{5}{1} \times \binom{4}{1} \times \binom{3}{1} \times \frac{5!}{2! 1! 1! 1!}.$$

Putting everything together, we obtain

$$P(A) = \frac{|A|}{|S|} = \frac{\binom{6}{1} \times \binom{5}{1} \times \binom{4}{1} \times \binom{3}{1} \times \frac{5!}{2! 1! 1! 1!}}{6^5}$$

Does the above solution seem correct? If not, then can you explain what is wrong?

Answer

It turns out the answer above is incorrect! How do we know this? Observe that

$$P(A) = \frac{|A|}{|S|} = \frac{\binom{6}{1} \times \binom{5}{1} \times \binom{4}{1} \times \binom{3}{1} \times \frac{5!}{2! 1! 1! 1!}}{6^5} \approx 2.78$$

and so clearly something is wrong since for any event A of sample space S , we proved that $P(A) \leq 1$.

We will now explore what is wrong in our solution and why.

Definition: Overcount

An **overcount** is when any distinct outcome has been incorrectly counted as multiple distinct outcomes.

The above example and definition raises two questions:

1. How can we recognize if an overcount has occurred?
2. If an overcount has occurred, then is there anything we can do to fix our answer?

We first answer question 1 by providing a recipe to check for overcounting.

Recipe to Check for an Overcount

1. After writing down the steps for computing the cardinality of some event, go through your steps and fix an actual outcome which we will denote by s_1 .
2. Now go through the steps again and change your choice on at least one step. Call the resulting outcome s_2 .
3. If for some change in choice, $s_1 = s_2$, then you have overcounted. If for all possible changes in your selection for s_2 , $s_1 \neq s_2$, then we have not overcounted.

Allow us to apply the recipe to our poker dice question discussed above. We do so by creating a chart:

Steps	s_1	s_2
1) We first choose a value of a . That is, we must select the denomination to be repeated. There are $\binom{6}{1}$ ways to do so.		
2) We now must select a value of b that is different from a . There are $\binom{5}{1}$ ways to do so.		
3) We now must select a value of c that is different from a and b . There are $\binom{4}{1}$ ways to do so.		
4) We now must select a value of d that is different from a , b and c . There are $\binom{3}{1}$ ways to do so.		
5) Finally, we must impose an ordering on a, a, b, c, d . There are $\frac{5!}{2!1!1!1!}$ ways to do so.		

We will now go through our steps and fix an outcome for s_1 . Let us suppose we come up with the following choices:

Steps	s_1	s_2
1) We first choose a value of a . That is, we must select the denomination to be repeated. There are $\binom{6}{1}$ ways to do so.	4, 4	
2) We now must select a value of b that is different from a . There are $\binom{5}{1}$ ways to do so.	3	
3) We now must select a value of c that is different from a and b . There are $\binom{4}{1}$ ways to do so.	1	
4) We now must select a value of d that is different from a , b and c . There are $\binom{3}{1}$ ways to do so.	2	
5) Finally, we must impose an ordering on a, a, b, c, d . There are $\frac{5!}{2!1!1!1!}$ ways to do so.	(4, 2, 1, 4, 3)	

Here, $s_1 = (4, 2, 1, 4, 3)$. To complete our recipe, we check if there is a way to change our choice on at least one step that will result in generating s_1 again. If there is a way to do so, then we have overcounted. Let us change our choice on steps 2, 3, and 4.

Steps	s_1	s_2
1) We first choose a value of a . That is, we must select the denomination to be repeated. There are $\binom{6}{1}$ ways to do so.	4, 4	4, 4
2) We now must select a value of b that is different from a . There are $\binom{5}{1}$ ways to do so.	3	1
3) We now must select a value of c that is different from a and b . There are $\binom{4}{1}$ ways to do so.	1	2
4) We now must select a value of d that is different from a, b and c . There are $\binom{3}{1}$ ways to do so.	2	3
5) Finally, we must impose an ordering on a, a, b, c, d . There are $\frac{5!}{2! 1! 1! 1!}$ ways to do so.	(4, 2, 1, 4, 3)	(4, 2, 1, 4, 3)

Here $s_2 = (4, 2, 1, 4, 3)$. Since $s_1 = s_2$, then we have overcounted. This leads to our second question. How do we fix an overcount?

📌 Method 1 to Fixing an Overcount

To fix an overcount, you may simply divide by the number of ways you have overcounted.

To use Method 1, we have to understand how did we overcount? Looking back at our steps, we see that the order in which we select b, c and d does not matter since their arrangement can yield the same outcome. How many ways can we arrange b, c and d ? There are $3!$ ways to do so. Hence, all we have to do is divide $|A|$ by $3!$ to obtain the correct answer.

$$P(A) = \frac{|A|}{|S|} = \frac{\left[\binom{6}{1} \binom{5}{1} \binom{4}{1} \binom{3}{1} \frac{5!}{2! 1! 1! 1!} \right] / 3!}{6^5} \approx 0.4630$$

📌 Method 2 to Fixing an Overcount

To fix an overcount, redo the counting process and condense the steps so that the overcount no longer occurs in the first place.

Consider the follow revision in our counting process:

Old Steps	New Steps
1) We first choose a value of a . That is, we must select the denomination to be repeated. There are $\binom{6}{1}$ ways to do so.	1) We first choose a value of a . That is, we must select the denomination to be repeated. There are $\binom{6}{1}$ ways to do so.
2) We now must select a value of b that is different from a . There are $\binom{5}{1}$ ways to do so.	2) We now select the values of b, c and d . There are $\binom{5}{3}$ ways to do so.
3) We now must select a value of c that is different from a and b . There are $\binom{4}{1}$ ways to do so.	

4) We now must select a value of d that is different from a , b and c . There are $\binom{3}{1}$ ways to do so.

5) Finally, we must impose an ordering on a, a, b, c, d . There are $\frac{5!}{2! 1! 1! 1!}$ ways to do so.

3) Finally, we must impose an ordering on a, a, b, c, d . There are $\frac{5!}{2! 1! 1! 1!}$ ways to do so.

$$P(A) = \frac{|A|}{|S|} = \frac{\binom{6}{1} \binom{5}{3} \frac{5!}{2! 1! 1! 1!}}{6^5} \approx 0.4630$$

Notice how steps 2, 3 and 4 under the old method have now been collapsed into one step in the new method.

4.2: Overcounting is shared under a [CC BY](#) license and was authored, remixed, and/or curated by LibreTexts.

CHAPTER OVERVIEW

Chapter 5: Week 5

[5.1: Review 1](#)

[5.2: Exam 1](#)

[Chapter 5: Week 5](#) is shared under a [CC BY](#) license and was authored, remixed, and/or curated by LibreTexts.

5.1: Review 1

5.1: Review 1 is shared under a [CC BY](#) license and was authored, remixed, and/or curated by LibreTexts.

5.2: Exam 1

5.2: Exam 1 is shared under a [CC BY](#) license and was authored, remixed, and/or curated by LibreTexts.

CHAPTER OVERVIEW

Chapter 6: Week 6

6.1: Conditional Probabilities

6.2: Independence

Chapter 6: Week 6 is shared under a [CC BY](#) license and was authored, remixed, and/or curated by LibreTexts.

6.1: Conditional Probabilities

The topic of conditional probabilities is one of the core concepts used in machine learning. As usual, we take a motivated approach towards this new topic and show how it relates to our theory that we have discussed so far.

In the last section, we formally proved that for any event E of a simple sample space S that

$$P(E) = \frac{|E|}{|S|}$$

Our goal is to now use this formula to motivate the answer to a rather deep question. Before we state this question allow us to be a little bit informal and abuse notation for the sake of intuition.

When we say for that a simple sample space S ,

$$P(E) = \frac{|E|}{|S|}$$

we are informally saying that

$$P(A) = \frac{|E|}{|S|} = \frac{\text{all the ``stuff'' in } E \text{ that can occur}}{\text{all the ``stuff'' which can possibly happen}}$$

In fact, even when you are not in a simple sample space, the same idea holds. Notice that

$$P(E) = \frac{P(E)}{P(S)} = \frac{\text{all the ``stuff'' in } E \text{ that can occur}}{\text{all the ``stuff'' which can possibly happen}}$$

With this idea in mind, let us consider the following question: have you ever played the MegaMillions before?

If not, then let us take a look at the following webpage: <https://www.megamillions.com/How-to-Play.aspx>

According to this webpage, the MegaMillions is played according to the following rules: there are two machines where Machine 1 contains 70 white balls, numbered 1 through 70, and Machine 2 contains 25 golden balls numbered 1 through 25. From Machine 1, five white balls are selected at random, one at a time, without replacement. From Machine 2, a single golden ball is selected. If a player is able to guess which five white balls are selected (in any order), and the player is able to guess which golden ball is selected, then they win the jackpot. According to the website, the probability that we win the jackpot (for a single ticket) is $\frac{1}{302,575,350}$. Is this number accurate? Let us find out.

✓ Example 1 6.1.1

There are two machines where Machine 1 contains 70 white balls, numbered 1 through 70, and Machine 2 contains 25 golden balls numbered 1 through 25. From Machine 1, five white balls are selected at random, one at a time, without replacement. From Machine 2, a single golden ball is selected. If a player is able to guess which five white balls are selected (in any order), and the player is able to guess which golden ball is selected, then they win the jackpot. Suppose you play a single ticket. What is the probability that you win the jackpot?

Solution

Let $A = \{\text{we win the jackpot}\} = \{\text{our lottery numbers were drawn}\}$.

Let S denote the set of all possible outcomes where an outcome is a subset of the 5 white balls drawn paired with a single golden ball.

Since the selection happens at random, S is a simple sample space.

$$|S| = \binom{70}{5} \times \binom{25}{1} = 302,575,350.$$

Since we only played a single ticket, $|A| = 1$.

$$\text{Hence } P(A) = \frac{1}{302,575,350}.$$

Hence the number reported on the webpage is indeed accurate. Observe that the above computation shows us that there is a small chance that we win the lottery. **Hence, initially, the probability we win the lottery is nonzero.**

For concreteness, let us suppose that on our lottery ticket, we said that the white balls numbered 24, 32, 55, 19 and 12 will be selected and the golden ball will be 17.

While watching the lottery drawing live, the first white ball that is drawn is numbered 1.

What is the probability that we win the jackpot now?

Answer

0.

The reason why is because 1 appears on the winning ticket. Since 1 does not appear on our ticket, then it means we did not win the lottery. Notice that **after learning some information, the probability we win the lottery is zero.**

Reflecting back on the situation we see that before $P(A) > 0$. However, after learning some new information that we will label by B , the $P(A \text{ GIVEN } B) = 0$. That is, the presence of new information affected the probability of A .

For this particular example, the effect that B had on the probability of A was obvious. However, for other situations, it may not be so clear on how to update/change/revise the probability of A . For these situations, we would like to develop a formula which tells us how the probability of A changes in the presence of new information B . And so the deep question that we will pose is: how can we update the the probability of A given that some other event, B , has occurred ?

Allow us to informally answer this deep question by abusing notation.

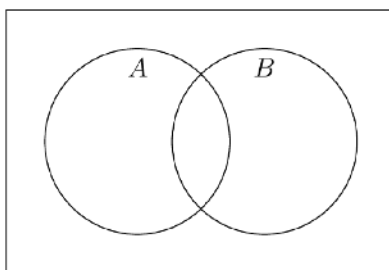
As mentioned above,

$$P(E) = \frac{P(E)}{P(S)} = \frac{\text{all the ``stuff'' in E that can occur}}{\text{all the ``stuff'' which can possibly happen}}$$

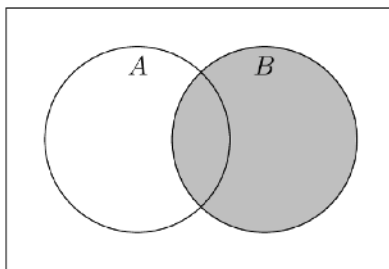
This means

$$P(A \text{ given } B) = \frac{\text{all the ``stuff'' in A that can occur}}{\text{all the ``stuff'' which can possibly happen}} \quad (6.1.1)$$

In order to identify the numerator and denominator, we will consider the following diagram:

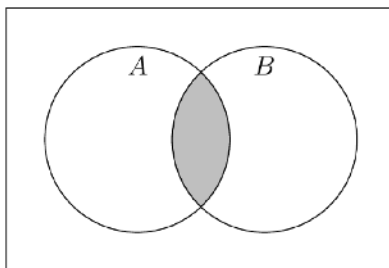


Let us first tackle the denominator of Equation 6.1.1. When we say $P(A \text{ given } B)$, we are saying that the event B has happened. That is, anything outside of B cannot occur. This means all the ``stuff'' which can possibly happen is precisely B .



Since B represents the set of all possibilities, B plays the role of our new (reduced) sample space.

We will now study the numerator of Equation 6.1.1. The numerator is all the "stuff" in A that can happen. Given that our new sample space is B , in order for A to occur, we must be in the following region:



That is, given that B has occurred, in order for A to occur, we must be in $A \cap B$. Putting everything together, we obtain the following:

$$P(A \text{ given } B) = \frac{\text{all the ``stuff'' in } A \text{ that can occur}}{\text{all the ``stuff'' which can possibly happen}} = \frac{P(A \cap B)}{P(B)} \quad (6.1.2)$$

Thus, we have answered the question of how the presence of new information B changes the probability of A and we take this to be our definition of a conditional probability.

Definition: The Conditional Probability

Suppose that we learn an event B has occurred and we wish to compute the probability of another event A taking into account the occurrence of B . The new probability of A is called the conditional probability of the event A given that the event B has occurred and is denoted by $P(A|B)$ which is read as "the probability of A given B ". If $P(B) > 0$, then we compute this probability as

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad (6.1.3)$$

✓ Example 2 6.1.2

Suppose two fair dice are rolled. Given that the first dice landed on a "3", what is the probability that a sum of 5 is obtained?

Answer

We are asked to find the probability that a sum of 5 is obtained after learning that the first dice landed on a "3". To answer this question, we will let $A = \{\text{the sum is a 5}\}$ and we will let $B = \{\text{the first dice is a 3}\}$. By our definition of conditional probability, we know that

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

To complete this problem, we need to find two probabilities. We must compute $P(A \cap B)$ and $P(B)$. Let us first tackle the denominator, $P(B)$.

The denominator is asking us to find the probability that the first dice lands on a 3. Recall that the experiment is that two fair dice are rolled. From the previous section, we recognize that the sample space of this experiment is a simple sample space. That is, the sample space can be thought of as the set of all outcomes where an outcome is an ordered pair. The first entry in the ordered pair denotes the outcome of rolling Die 1 and the second entry in the ordered pair denotes the outcome of rolling Die 2. Notice that S is a simple sample space because there is no reason to believe that a certain ordered pair is more likely than another ordered pair since the dice are fair. Visually we can list out the outcomes in S via the following chart:

	Die 2	1	2	3	4	5	6
Die 1							

	Die 2	1	2	3	4	5	6
1		(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
2		(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
3		(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
4		(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
5		(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
6		(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

Hence $P(B) = \frac{|B|}{|S|} = \frac{6}{36}$.

Similarly, the numerator, $P(A \cap B) = \frac{|A \cap B|}{|S|}$.

To find $|A \cap B|$, we ask ourselves how many outcomes belong to $A \cap B$? That is, how many outcomes correspond to a sum of 5 where the first die lands on a 3? The only such outcome is (3,2). Hence, $P(A \cap B) = \frac{|A \cap B|}{|S|} = \frac{1}{36}$.

Putting everything together yields

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{|A \cap B|}{|S|}}{\frac{|B|}{|S|}} = \frac{\frac{1}{36}}{\frac{6}{36}} = \frac{1}{6}$$

While this completes the problem, notice that before, $P(A) = 4/36$. After learning the event B has taken place, $P(A|B) = \frac{1}{6} = \frac{6}{36}$. Thus, learning that B has occurred has made A more likely to occur.

✓ Example 2 6.1.3

Suppose two fair dice are rolled. Given that the sum of the two dice is 5, what is the probability that the first dice lands on a 3?

Answer

For convenience, let us stick to the notation that we used in Example 1. We will let $A = \{\text{the sum is a 5}\}$ and $B = \{\text{the first dice is a 3}\}$. We are now asked to find $P(B|A)$. By a similar argument to the above problem,

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{\frac{|B \cap A|}{|S|}}{\frac{|A|}{|S|}} = \frac{\frac{1}{36}}{\frac{4}{36}} = \frac{1}{4}$$

While this completes the problem, notice that before, $P(B) = \frac{6}{36} = \frac{1}{6}$. After learning the event A has taken place, $P(B|A) = \frac{1}{4} = \frac{9}{36}$. Thus, learning that A has occurred has made B more likely to occur.

Before we consider the next example, allow us to make an observation. In the lottery example, we saw that $P(A) > P(A|B)$. In the example above, we saw that $P(A) < P(A|B)$.

Hence, the presence of new information B can either increase the probability of A (as in the dice example) or decrease the probability of A (as in the lottery example). However, there is a third case we may wish to consider. Namely, is it possible that some new information B can have no affect on the probability of A ? Allow us to consider the following example which will answer this question. (Note: You should highlight this problem since it will be the central focus in the next section!).

✓ Example 6.1.4

Suppose a card is selected at random from a regular deck of cards. Let $E = \{\text{the selected card is an ace}\}$ and let $F = \{\text{the selected card is a spade}\}$. Find $P(E)$, $P(F)$, $P(E|F)$ and $P(F|E)$.

Answer

Since the card is selected at random, we are in a simple sample space. Hence $P(E) = \frac{|E|}{|S|} = \frac{4}{52} = \frac{1}{13}$. By a similar argument, $P(F) = \frac{|F|}{|S|} = \frac{13}{52} = \frac{1}{4}$.

To find $P(E|F)$, we simply note that

$$P(E|F) = \frac{P(E \cap F)}{P(F)} = \frac{\frac{|E \cap F|}{|S|}}{\frac{|F|}{|S|}} = \frac{\frac{1}{52}}{\frac{13}{52}} = \frac{1}{13}$$

Notice that $P(E) = P(E|F)$.

Meanwhile, to find $P(F|E)$, we simply note that

$$P(F|E) = \frac{P(F \cap E)}{P(E)} = \frac{\frac{|F \cap E|}{|S|}}{\frac{|E|}{|S|}} = \frac{\frac{1}{52}}{\frac{4}{52}} = \frac{1}{4}$$

Notice that $P(F) = P(F|E)$.

Allow us to make a comment regarding the above problem as to why it is trivial that $P(E) = P(E|F)$. Suppose that we were playing a game where you draw a card at random and if that card is an ace, then you would win the game. Before you draw your card, the probability that you get an ace is $\frac{4}{52} = \frac{1}{13}$.

Suppose after you selected a face down card, I take a quick peek at the card and inform you that the card you drew was a spade. Does this new knowledge about the selected card affect the probability of you winning? Not at all! You see, with the knowledge that the card is a spade, you know that the card has thirteen possibilities. Of these thirteen possibilities, only one of them is an ace and since all of these outcomes are equally likely, the probability that the card is an ace is $\frac{1}{13}$ which is precisely our initial probability.

Note that we can perform a similar analysis to explain why $P(F) = P(F|E)$.

🔗 Theorem 6.1.5

Theorem: Suppose B is an event of sample space S such that $P(B) > 0$. Then for any event A in S ,

$$P(A^c|B) = 1 - P(A|B)$$

Proof:

Let as an exercise to the reader.

🔗 Theorem: The Basic Conditioning Rule 6.1.6

Theorem: Suppose A and B are events in a sample space S .

1. If $P(A) > 0$, then $P(A \cap B) = P(A)P(B|A)$.
2. If $P(B) > 0$, then $P(A \cap B) = P(B)P(A|B)$.

Proof:

We only show the proof of (1) as the proof of (2) is similar.

Consider the expression $P(A)P(B|A)$. Substituting the definition of $P(B|A)$ into $P(A)P(B|A)$ yields the following:

$$\begin{aligned} P(A)P(B|A) &= P(A) \times \frac{P(B \cap A)}{P(A)} \\ &= P(B \cap A) \\ &= P(A \cap B) \end{aligned}$$

Notice the above result tells us that we have a choice to make when studying an intersection. Whenever we are dealing with $P(A \cap B)$, we must decide if we wish to rewrite this as $P(A)P(B|A)$ or as $P(B)P(A|B)$. How do we know which one to select? The answer is that it will depend either on the experiment or the probabilities available to us within the problem.

✓ Example 6.1.7

Suppose an urn has 10 red marbles and 6 blue marbles. An experiment consists of drawing two marbles from the bowl, one at a time without replacement. If we assume at each draw, each marble is equally likely to be chosen, then find the probability that both marbles are red. (Try this problem on your own before reading our solution!).

Answer

We will answer this question in two different ways.

Solution 1: In our first solution, we stick to the counting framework introduced in Section 4.1. Let $A = \{\text{both marbles are red}\}$. Let S denote the set of all outcomes where an outcome is a subset of two marbles. S is a simple sample space since there is no reason to believe that a certain subset of two marbles is more likely than another subset of two marbles. $|S| = \binom{16}{2}$ while $|A| = \binom{10}{2}$. Putting everything together, we see that

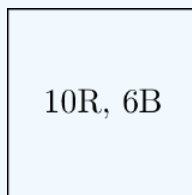
$$P(A) = \frac{|A|}{|S|} = \frac{\binom{10}{2}}{\binom{16}{2}} = 0.375$$

Solution 2: Recall from Section 2.1 that we said in rephrasing an event, it may be helpful to ask ourselves "can we describe this event by using the words *or/and*?"

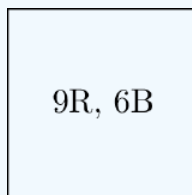
For us, $A = \{\text{both marbles are red}\}$ can be rephrased as $A = \{\text{marble 1 is red and marble 2 is red}\}$. In order to introduce suitable notation to describe A , we define $R_i = \{\text{marble } i \text{ is red}\}$. Thus, $A = R_1 \cap R_2$. This gives us the following:

$$P(A) = P(R_1 \cap R_2) = P(R_1)P(R_2|R_1)$$

Note that our starting configuration looks like this:



Clearly, $P(R_1) = \frac{10}{16}$. To find $P(R_2|R_1)$, we note that given R_1 , our new bowl looks like this:



And so given that R_1 has occurred, we now face the above composition. Hence the probability of R_2 with respect to our new sample space is given by $\frac{9}{15}$. Putting everything together, we see that

$$P(A) = P(R_1 \cap R_2) = P(R_1)P(R_2|R_1) = \frac{10}{16} \times \frac{9}{15} = 0.375$$

The Basic Conditioning Rule tells us how to deal with the probability of an intersection of two events. However, there is no reason to limit ourselves to studying the intersection of only two events. Instead, we may wish to deal with the intersection of three or four or in general, k events. To do so, we use The Generalized Conditioning Rule.

Theorem: The Generalized Conditioning Rule 6.1.8

Theorem: Suppose that $A_1, A_2, A_3, \dots, A_n$ are events in a sample space S such that $P(A_1 \cap A_2 \cap A_3 \cap \dots \cap A_{n-1}) > 0$. Then

$$P(P(A_1 \cap A_2 \cap A_3 \cap \dots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \dots P(A_n|A_1 \cap A_2 \cap \dots \cap A_{n-1}) \quad (6.1.4)$$

Proof

Let as an exercise to the reader.

✓ Example 6.1.9

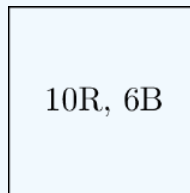
Suppose an urn has 10 red marbles and 6 blue marbles. An experiment consists of drawing three marbles from the bowl, one at a time without replacement. If we assume at each draw, each marble is equally likely to be chosen, then find the probability that the first two marbles are red and the third marble is blue.

Answer

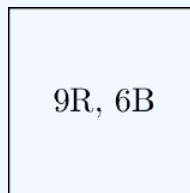
Let $A = \{\text{the first two marbles are red and the third marble is blue}\}$. We can rephrase this event as $A = \{\text{marble 1 is red and marble 2 is red and marble 3 is blue}\}$. In order to introduce suitable notation to describe A , we define $R_i = \{\text{marble } i \text{ is red}\}$ and $B_i = \{\text{marble } i \text{ is blue}\}$. Thus, $A = R_1 \cap R_2 \cap B_3$. This gives us the following:

$$P(A) = P(R_1 \cap R_2 \cap B_3) = P(R_1)P(R_2|R_1)P(B_3|R_1 \cap R_2)$$

Note that our starting configuration looks like this:



Clearly, $P(R_1) = \frac{10}{16}$. To find $P(R_2|R_1)$, we note that given R_1 , our new bowl looks like this:



And so given that R_1 has occurred, we now face the above composition. Hence the probability of R_2 with respect to our new sample space is given by $\frac{9}{15}$.

Similarly, to find $P(B_3|R_1 \cap R_2)$ we note that given $R_1 \cap R_2$, our new bowl looks like this:

8R, 6B

And so given that $R_1 \cap R_2$, has occurred, we now face the above composition. Hence the probability of B_3 with respect to the newest sample space is $\frac{6}{14}$.

Putting everything together, we see that

$$P(A) = P(R_1 \cap R_2 \cap B_3) = P(R_1)P(R_2|R_1)P(B_3|R_1 \cap R_2) = \frac{10}{16} \times \frac{9}{15} \times \frac{6}{14} \approx 0.1607$$

✓ Example 6.1.10

Suppose an urn has 10 red marbles and 6 blue marbles. An experiment consists of drawing three marbles from the bowl, one at a time. Suppose each time a marble is selected from the bowl, the selected marble is replaced along with two additional marbles of the same color. If we assume at each draw, each marble is equally likely to be chosen, then find the probability that the first two marbles are red and the third marble is blue.

Answer

Let $A = \{\text{the first two marbles are red and the third marble is blue}\}$. We can rephrase this event as $A = \{\text{marble 1 is red and marble 2 is red and marble 3 is blue}\}$. In order to introduce suitable notation to describe A , we define $R_i = \{\text{marble } i \text{ is red}\}$ and $B_i = \{\text{marble } i \text{ is blue}\}$. Thus, $A = R_1 \cap R_2 \cap B_3$. This gives us the following:

$$P(A) = P(R_1 \cap R_2 \cap B_3) = P(R_1)P(R_2|R_1)P(B_3|R_1 \cap R_2)$$

Note that our starting configuration looks like this:

10R, 6B

Clearly, $P(R_1) = \frac{10}{16}$. To find $P(R_2|R_1)$, we note that given R_1 , our new bowl looks like this:

12R, 6B

And so given that R_1 has occurred, we now face the above composition. Hence the probability of R_2 with respect to our new sample space is given by $\frac{12}{18}$.

Similarly, to find $P(B_3|R_1 \cap R_2)$ we note that given $R_1 \cap R_2$, our new bowl looks like this:

14R, 6B

And so given that $R_1 \cap R_2$, has occurred, we now face the above composition. Hence the probability of B_3 with respect to the newest sample space is $\frac{6}{20}$.

Putting everything together, we see that

$$P(A) = P(R_1 \cap R_2 \cap B_3) = P(R_1)P(R_2|R_1)P(B_3|R_1 \cap R_2) = \frac{10}{16} \times \frac{12}{18} \times \frac{6}{20} \approx 0.125$$

6.1: Conditional Probabilities is shared under a [CC BY](#) license and was authored, remixed, and/or curated by LibreTexts.

6.2: Independence

In the previous section, we saw how information about some event B can affect the probability of some other event, A . In the initial lottery example, we saw that learning some information B decreased the probability of A . In the dice example, we saw that learning some information B increased the probability of A . But then we saw the following interesting example:

✓ Example 1 6.2.1

Suppose a card is selected at random from a regular deck of cards. Let $E = \{\text{the selected card is an ace}\}$ and let $F = \{\text{the selected card is a spade}\}$. Find $P(E)$, $P(F)$, $P(E|F)$ and $P(F|E)$.

Answer

Since the card is selected at random, we are in a simple sample space. Hence $P(E) = \frac{|E|}{|S|} = \frac{4}{52} = \frac{1}{13}$. By a similar argument, $P(F) = \frac{|F|}{|S|} = \frac{13}{52} = \frac{1}{4}$.

To find $P(E|F)$, we simply note that

$$P(E|F) = \frac{P(E \cap F)}{P(F)} = \frac{\frac{|E \cap F|}{|S|}}{\frac{|F|}{|S|}} = \frac{\frac{1}{52}}{\frac{13}{52}} = \frac{1}{13}$$

Notice that $P(E) = P(E|F)$.

Meanwhile, to find $P(F|E)$, we simply note that

$$P(F|E) = \frac{P(F \cap E)}{P(E)} = \frac{\frac{|F \cap E|}{|S|}}{\frac{|E|}{|S|}} = \frac{\frac{1}{52}}{\frac{4}{52}} = \frac{1}{4}$$

Notice that $P(F) = P(F|E)$.

We noted that $P(E) = P(E|F)$ and $P(F) = P(F|E)$. Allow us to now interpret this.

When we say $P(E) = P(E|F)$, we are saying that *the occurrence of F has not affected the probability of E* .

Similarly, when we say $P(F) = P(F|E)$, we are saying that *the occurrence of E has not affected the probability of F* .

Hence, we might be inclined to say that the probability of the event E **does not depend** upon the knowledge of F occurring and the probability of the event F **does not depend** upon the knowledge of E occurring.

And finally, the last leap for us to make is that a simple way of saying "**does not depend**" is to say that these events are **independent**. That is, E and F are independent events since the occurrence of one event does not affect the probability of the other event, and vice-versa.

When we use the word "**independent**" in the above paragraph, we do so colloquially. Allow us to mathematically define what it means for two events to be independent.

Definition: Independence of Two Events

Definition: Two events, E and F are said to be **independent** provided that

$$P(E \cap F) = P(E)P(F) \quad (6.2.1)$$

If $P(E \cap F) \neq P(E)P(F)$, then we say that E and F are **dependent**.

In words, the formal definition is saying that for two events to be independent, it means that the probability of their intersection is equal to the product of the individual probabilities. Note that the formal definition of independence gives us our intuitive idea of what independence should be. To see this, note that by using the formal definition, we obtain the following two equalities:

$$\begin{aligned} P(E|F) &\stackrel{\text{def.}}{=} \frac{P(E \cap F)}{P(F)} \stackrel{\text{def.}}{=} \frac{P(E)P(F)}{P(F)} = P(E) \\ P(F|E) &\stackrel{\text{def.}}{=} \frac{P(F \cap E)}{P(E)} \stackrel{\text{def.}}{=} \frac{P(F)P(E)}{P(E)} = P(F) \end{aligned}$$

✓ Example 2.6.2.2

Suppose that two printers are operated independently of each other. Further suppose that the probability that Printer 1 will malfunction in a 24 hour period is $\frac{1}{3}$ and the probability that Printer 2 will malfunction in the same 24 hour period is $\frac{1}{4}$. Find the probability that at least one machine will malfunction during the same 24-hour period.

Answer

Let $A = \{\text{at least one printer will malfunction during the same 24-hour period}\}$. Notice that we can rewrite A as

$$\begin{aligned} A &= \{\text{at least one printer will malfunction during the same 24-hour period}\} \\ &= \{\text{Printer 1 will malfunction OR Printer 2 will malfunction}\} \\ &= \{\text{Printer 1 will malfunction}\} \cup \{\text{Printer 2 will malfunction}\} \\ &= M_1 \cup M_2 \end{aligned}$$

where $M_i = \{\text{Machine } i \text{ will malfunction during the 24-hour period}\}$. Hence we obtain the following:

$$\begin{aligned} P(A) &= P(M_1 \cup M_2) \\ &= P(M_1) + P(M_2) - \underbrace{P(M_1 \cap M_2)}_{\text{independence}} \\ &= P(M_1) + P(M_2) - P(M_1)P(M_2) \\ &= \frac{1}{3} + \frac{1}{4} - \left(\frac{1}{3} \times \frac{1}{4}\right) \\ &= 0.5 \end{aligned}$$

⚠ Caution

When asked to explain what it means for two events to be independent, some students falsely paraphrase the idea and state something along the lines of *independence means that one event does not affect the other event*. With this type of thinking, they are then often led to believe that since one event does not affect the other event, then independence means that these events must be separate or mutually exclusive. However, this is incorrect! Can you explain why?

Answer

Independence means that one event does not affect *the probability* of the other event and vice-versa. However, it is possible that two events are independent and one event *does affect the outcomes* in the other event.

To see this, let us think back to the card example considered in the beginning of this section. Notice that $P(E) = P(E|F)$ and $P(F) = P(F|E)$ and so one event did not affect the probability of the other event and vice-versa. However, notice that the knowledge of E does affect the outcomes which can happen in F and vice-versa. You see, before we learn anything, the possible outcomes in E are $A\heartsuit, A\clubsuit, A\diamondsuit, A\spadesuit$. After learning that F has occurred, the only way for E to occur is to obtain $A\spadesuit$. That is, the other outcomes in E can no longer occur and so learning that the event F has occurred has affected which outcomes in E can occur. Despite this, the overall probability of E remains unchanged.

In summary, if two events are independent, then each event does not affect the probability of the other's but it is possible that each event affects which outcomes can occur in the other.

Additionally, there is an important distinction to be made. Notice that if E and F are two events with nonzero probabilities, then it is impossible for E and F to be both mutually exclusive and independent. To see why, let us assume that it is possible to have two events E and F where $P(E) > 0$ and $P(F) > 0$ such that E and F are both independent and disjoint. Independence guarantees that

$$\underbrace{P(E \cap F) = P(E)P(F)}_{\text{independence}} > 0$$

while disjoint guarantees

$$\underbrace{P(E \cap F) = P(\emptyset) = 0}_{\text{mutually exclusive}}$$

Hence $P(E \cap F) > 0$ and $P(E \cap F) = 0$. Contradiction!

We summarize the above discussion in the following remark.

Remark

In summary, the independence of two events means that the occurrence of one event does not affect the probability of the other event and vice-versa.*

Additionally, if you have two independent events, both of which have a non-zero probability of occurring, then it is impossible for the events to also be disjoint.

*Note that we will be revising this statement in a few moments.

Our next theorem will be especially important for later examples.

Theorem 6.2.3

Theorem: If E and F are independent events, then the following events are also independent:

1. E^c and F
2. E and F^c
3. E^c and F^c

Answer

We discuss the proof of (1) and leave the other two proofs as an exercise for homework.

Let us assume that E and F are independent events. We must show that E^c and F are independent. That is, we must show $P(E^c \cap F) = P(E^c)P(F)$.

$$\begin{aligned} P(E^c \cap F) &= P(F) - \underbrace{P(E \cap F)}_{\text{independence}} \\ &= P(F) - P(E)P(F) \\ &= P(F)[1 - P(E)] \\ &= P(F)P(E^c) \\ &= P(E^c)P(F) \end{aligned}$$

Intuition of Independence (Revised)

By virtue of the above theorem, we see that if E and F are independent then:

1. $P(E) = P(E|F) = P(E|F^c)$ and
2. $P(F) = P(F|E) = P(F|E^c)$.

Thus, we can say that intuitively, independence means that the occurrence *or nonoccurrence* of one event does not affect the probability of the other event and vice-versa.

Our next step is to generalize the definition of independence for more than two events. Most students reason that since for two events to be independent, it means that the probability of their intersection is equal to the product of the individual probabilities, then the same must be true for three events to be independent. That is, most students will guess that three events, E, F and G are independent provided that $P(E \cap F \cap G) = P(E)P(F)P(G)$.

However, there is more to it! Since E, F and G are independent then every subcollection of events is also independent. That is, E and F are independent, E and G are independent and F and G are independent. Hence we have the following definition:

Definition: Independence of Three Events

Definition: Three events, E, F and G are said to be **mutually independent** provided that all of the following equations are satisfied:

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

The following definition generalizes the above idea in order to extend independence to k events.

Definition: General Independence

Definition: The k events, $A_1, A_2, \dots, A_k$ are **mutually independent** provided that for every subset $A_{i_1}, A_{i_2}, \dots, A_{i_j}$ of j of these events, ($j = 2, 3, 4, \dots, k$),

$$P(A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_j}) = P(A_{i_1}) \times P(A_{i_2}) \times \dots \times P(A_{i_j})$$

✓ Example 3 6.2.4

In reliability theory, a parallel system refers to a system which will function as long as at least one of its components are working. Suppose the batteries in the Mars Rover is a parallel system that has three components. Further suppose that the probability that any component is working is 0.99. If the components operate independently of each other, then what is the probability that the system will function?

Solution 1

We will first answer this question directly. Let $E = \{\text{the system will function}\}$. As usual, we pause for a moment and ask if we can re-write E . We notice we can write

$$\begin{aligned} E &= \{\text{the system will function}\} \\ &= \{\text{at least one component is working}\} \\ &= \{\text{Component 1 works OR Component 2 works OR Component 3 works}\} \\ &= C_1 \cup C_2 \cup C_3 \end{aligned}$$

where $C_i = \{\text{Component } i \text{ is working}\}$. By Equation 1.3.13, we obtain the following:

$$\begin{aligned} P(E) &= P(C_1 \cup C_2 \cup C_3) \\ &= P(C_1) + P(C_2) + P(C_3) - \underbrace{P(C_1 \cap C_2)}_{\text{indep.}} - \underbrace{P(C_1 \cap C_3)}_{\text{indep.}} - \underbrace{P(C_2 \cap C_3)}_{\text{indep.}} + \underbrace{P(C_1 \cap C_2 \cap C_3)}_{\text{indep.}} \\ &= P(C_1) + P(C_2) + P(C_3) - P(C_1)P(C_2) - P(C_1)P(C_3) - P(C_2)P(C_3) + P(C_1)P(C_2)P(C_3) \\ &= 0.99 + 0.99 + 0.99 - 0.99(0.99) - 0.99(0.99) - 0.99(0.99) + 0.99(0.99)(0.99) \\ &= 0.999999 \end{aligned}$$

Notice that if the Rover only had one battery component, then the Rover will function with a 99% chance. However, by adding an additional two components as part of a parallel system, the Rover will now have a higher chance of functioning correctly.

✓ Example 4 6.2.5

In reliability theory, a parallel system refers to a system which will function as long as at least one of its components are working. Suppose the batteries in the Mars Rover is a parallel system that has three components. Further suppose that the probability that any component is working is 0.99. If the components operate independently of each other, then what is the probability that the system will function?

Solution 2

We will now answer this question indirectly via the complement. Since $E = \{\text{the system will function}\} = \{\text{at least one component is working}\}$, then

$$\begin{aligned} E^c &= \{\text{none of the components are working}\} \\ &= \{\text{Component 1 is not working AND Component 2 is not working AND component 3 is not working}\} \\ &= \{\text{Component 1 is not working}\} \cap \{\text{Component 2 is not working}\} \cap \{\text{Component 3 is not working}\} \\ &= C_1^c \cap C_2^c \cap C_3^c \end{aligned}$$

Putting everything together yields

$$\begin{aligned} P(E) &= 1 - P(E^c) = 1 - \underbrace{P(C_1^c \cap C_2^c \cap C_3^c)}_{\text{indep.}} \\ &= 1 - P(C_1^c)P(C_2^c)P(C_3^c) \\ &= 1 - 0.01(0.01)(0.01) \\ &= 0.999999 \end{aligned}$$

✓ Example 5 6.2.6

Suppose an experiment consists of repeatedly flipping a fair coin until a heads appears for the first time at which point the experiment stops. Assuming that the flips of the coin are independent of each other, find the probability that the experiment will eventually terminate. That is, find the probability that a heads is eventually obtained.

Answer

Before we answer this, allow us to guess the solution. If we are given a fair coin and we are asked what is the probability that we will eventually obtain a heads, our intuition might tell us that the probability should be 1. That is, we are saying that given an infinite amount of coin flips, we are certain that eventually we will obtain a heads. Thus, we expect the answer to be 1. Allow us to now formally answer the problem.

We may answer this question directly as in Example 3 but we will opt for an indirect approach as in Example 4. (The direct approach is left as an exercise for homework). Let $E = \{\text{we eventually obtain our first heads}\}$. Then

$$\begin{aligned} E^c &= \{\text{it is not the case we eventually obtain our first heads}\} \\ &= \{\text{we never obtain our first heads}\} \\ &= \{\text{we always obtain a tails}\} \\ &= \{\text{flip 1 results in tails AND flip 2 results in tails AND flip 3 results in tails AND } \dots\} \\ &= T_1 \cap T_2 \cap T_3 \cap \dots \\ &= \end{aligned}$$

where $T_i = \{\text{we obtain a tails on flip } i\}$. Putting everything together yields

$$\begin{aligned} P(E) &= 1 - P(E^c) = 1 - \underbrace{P(T_1 \cap T_2 \cap T_3 \cap \dots)}_{\text{indep.}} \\ &= 1 - P(T_1)P(T_2)P(T_3) \dots \\ &= 1 - 0.5(0.5)(0.5) \dots \\ &= 1 - \lim_{n \rightarrow \infty} 0.5^n \\ &= 1 - 0 = 1 \end{aligned}$$

✓ Example 6 6.2.7

A sequence of five independent trials are to be performed. Each trial consists of administering a drug to a patient and then seeing if the drug is a success (meaning beneficial) for that patient or if the drug is a failure (meaning non-beneficial) for that patient. The probability the drug is successful for any patient is 0.75. Find the probability that the drug is successful for at least one patient.

Answer

We may answer this question directly as in Example 3, but we will opt for an indirect approach as in Examples 4 and 5. (The direct approach is left as an exercise for homework). Let $E = \{\text{the drug is successful for at least one patient}\}$. Then

$$\begin{aligned} E^c &= \{\text{the drug is successful for no patients}\} \\ &= \{\text{the drug is a failure for all patients}\} \\ &= \{\text{the drug is a failure for patient 1 AND the drug is a failure for patient 2 AND } \dots \text{ AND the drug is a failure for patient 5}\} \\ &= F_1 \cap F_2 \cap F_3 \cap F_4 \cap F_5 \end{aligned}$$

where $F_i = \{\text{the drug is a failure for patient } i\}$. Putting everything together yields

$$\begin{aligned} P(E) &= 1 - P(E^c) = 1 - \underbrace{P(F_1 \cap F_2 \cap F_3 \cap F_4 \cap F_5)}_{\text{indep.}} \\ &= 1 - P(F_1)P(F_2)P(F_3)P(F_4)P(F_5) \\ &= 1 - 0.25(0.25)(0.25)(0.25)(0.25) \dots \\ &= 1 - 0.25^5 \\ &= \approx 0.9990 \end{aligned}$$

Our last example for this section is particularly important as it's generalization will yield an important result.

✓ Example 7 6.2.8

A sequence of five independent trials are to be performed. Each trial consists of administering a drug to a patient and then seeing if the drug is a success (meaning beneficial) for that patient or if the drug is a failure (meaning non-beneficial) for that patient. The probability the drug is successful for any patient is 0.75. Find the probability that exactly three successes occur.

Answer

Let $E = \{\text{we obtain exactly three successes}\}$. Notice that the complement of E is that we obtain zero or one or two or four or five successes. Since the complement is more tedious than E itself, we will stick to answering this question directly.

As usual, we ask ourselves *is it possible to rewrite E ?* If we let $F_i = \{\text{the drug is a failure for patient } i\}$ and $S_i = \{\text{the drug is a success for patient } i\}$ then E is the union of all possible sequences of five trials where three of the trials are a success. That is,

$$E = (S_1 \cap S_2 \cap S_3 \cap F_4 \cap F_5) \cup (S_1 \cap S_2 \cap F_3 \cap F_4 \cap S_5) \cup \dots \cup (F_1 \cup S_2 \cup S_3 \cup S_4 \cap F_5)$$

Notice the events, $(S_1 \cap S_2 \cap S_3 \cap F_4 \cap F_5), (S_1 \cap S_2 \cap F_3 \cap F_4 \cap S_5), \dots, (F_1 \cup S_2 \cup S_3 \cup S_4 \cap F_5)$ are all disjoint and so

$$P(E) = P(S_1 \cap S_2 \cap S_3 \cap F_4 \cap F_5) + P(S_1 \cap S_2 \cap F_3 \cap F_4 \cap S_5) + \dots + P(F_1 \cup S_2 \cup S_3 \cup S_4 \cap F_5)$$

Notice that by independence, $P(S_1 \cap S_2 \cap S_3 \cap F_4 \cap F_5) = P(S_1)P(S_2)P(S_3)P(F_4)P(F_5) = 0.75^3 \times 0.25^2$ and similarly, $P(S_1 \cap S_2 \cap F_3 \cap F_4 \cap S_5) = 0.75^3 \times 0.25^2$ and $P(F_1 \cup S_2 \cup S_3 \cup S_4 \cap F_5) = 0.75^3 \times 0.25^2$. That is, since all of the events in the union of E have three successes and two failures, each event occurs with probability $0.75^3 \times 0.25^2$.

Hence,

$$\begin{aligned} P(E) &= P(S_1 \cap S_2 \cap S_3 \cap F_4 \cap F_5) + P(S_1 \cap S_2 \cap F_3 \cap F_4 \cap S_5) + \dots + P(F_1 \cup S_2 \cup S_3 \cup S_4 \cap F_5) \\ &= \underbrace{0.75^3 \times 0.25^2 + 0.75^3 \times 0.25^2 + \dots + 0.75^3 \times 0.25^2}_{\text{how many times are these added together?}} \end{aligned}$$

The number of times we have $0.75^3 \times 0.25^2$ is equal to the number of times we are able to have three successes among the five placeholders. That is, there are $\binom{5}{3} = 10$ possible arrangements. Putting everything together yields

$$\begin{aligned} P(E) &= P(S_1 \cap S_2 \cap S_3 \cap F_4 \cap F_5) + P(S_1 \cap S_2 \cap F_3 \cap F_4 \cap S_5) + \dots + P(F_1 \cup S_2 \cup S_3 \cup S_4 \cap F_5) \\ &= \underbrace{0.75^3 \times 0.25^2 + 0.75^3 \times 0.25^2 + \dots + 0.75^3 \times 0.25^2}_{\binom{5}{3}} \\ &= \binom{5}{3} 0.75^3 0.25^2 \end{aligned}$$

The above example can be generalized to show the following result:

Theorem 6.2.9

Theorem: Suppose a sequence of n independent trials are to be performed where each trial may result in either a success with probability p or may result in a failure with probability $1 - p$. The probability that we obtain exactly x successes (where x is either $0, 1, 2, \dots, n$) is given by

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

Proof:

The proof is left as a homework exercise. Hint: Generalize the above problem!

6.2: Independence is shared under a [CC BY](https://creativecommons.org/licenses/by/4.0/) license and was authored, remixed, and/or curated by LibreTexts.

CHAPTER OVERVIEW

Chapter 7: Week 7

[7.1: Bayes' Theorem](#)

[Chapter 7: Week 7](#) is shared under a [CC BY](#) license and was authored, remixed, and/or curated by LibreTexts.

7.1: Bayes' Theorem

In this section, we will explore an extremely powerful result which is called Bayes' Theorem. This result is so important that the core idea has been generalized and a whole course can be dedicated towards it. Before we state Bayes' Theorem allow us to note why Bayes' Theorem is important.

1. NVIDIA is a company which specializes in hardware and software, especially those relating to artificial intelligence. According to NVIDIA, "NB [Naive Bayes] algorithms have been shown to work well on text classification use cases. They are often applied to tasks such as filtering spam emails; predicting categories and sentiment for tweets, web pages, blog posts, user ratings, and forum posts; or ranking documents and web pages". More information can be found here: <https://developer.nvidia.com/blog/faster-text-classification-with-naive-bayes-and-gpus/>
2. An excellent resource to read in your spare time is *The Theory That Would Not Die (How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy)*. <https://yalebooks.yale.edu/book/9780300188226/the-theory-that-would-not-die/>

In order to understand Bayes' Theorem, we must recall a few definitions.

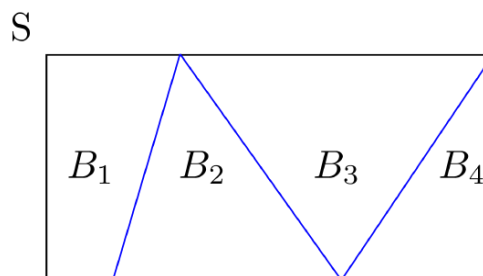
 **Definition: Mutually Exclusive Events, Exhaustive Events, Forms a Partition on S**

1. The events $E_1, E_2, \dots, E_k$ are said to be **mutually exclusive events** if $E_i \cap E_j = \emptyset$ for all $i \neq j$.
2. The events $E_1, E_2, \dots, E_k$ are said to be **exhaustive on S** if $E_1 \cup E_2 \cup \dots \cup E_k = S$.
3. The events $E_1, E_2, \dots, E_k$ are said to **form a partition on S** if $E_1, E_2, \dots, E_k$ are mutually exclusive events and exhaustive on S .

 **Example 1 7.1.1**

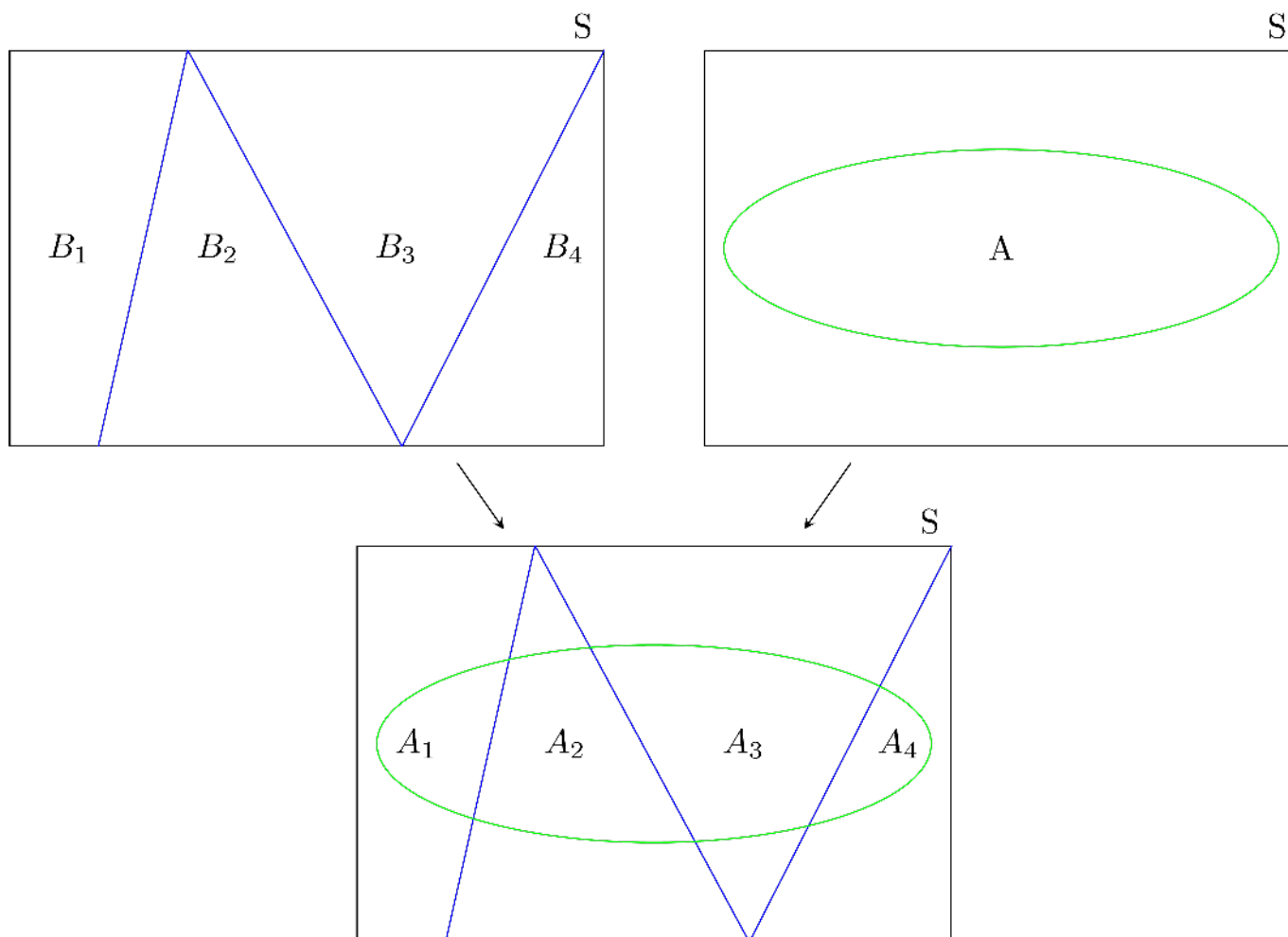
1. If $S = \{1, 2, 3, 4, 5, 6\}$ and $E_1 = \{1, 2, 3\}$ and $E_2 = \{4, 5, 6\}$ then E_1 and E_2 form a partition on S .
2. If $S = \{1, 2, 3, 4, 5, 6\}$ and $E_1 = \{1, 2\}$ and $E_2 = \{3, 4\}$ and $E_3 = \{5, 6\}$ then E_1, E_2, E_3 forms a partition on S .
3. If S is the sample space for some experiment, then A and A^c forms a partition on S .

At this point, we might wonder what exactly a partition does to a sample space. The answer is that a partition splits or fragments the sample space into disjoint pieces. For instance, if B_1, B_2, B_3 and B_4 partitions S , then the picture might look like this:



Notice that B_1, B_2, B_3 and B_4 are all mutually exclusive and that their union gives back the entire sample space S .

The question we will not consider is why are partitions important? Quite simply, partitions are important because they will help us compute the probability of an event A . How so? Well let us assume that B_1, B_2, B_3 and B_4 forms a partition on S and suppose we have some event A in S . Consider the following diagram:



We see that the event $A = A_1 \cup A_2 \cup A_3 \cup A_4 = (B_1 \cap A) \cup (B_2 \cap A) \cup (B_3 \cap A) \cup (B_4 \cap A)$. Notice that $(B_1 \cap A)$, $(B_2 \cap A)$, $(B_3 \cap A)$ and $(B_4 \cap A)$ are all mutually exclusive. This means we have the following:

$$\begin{aligned}
 P(A) &= P(A_1 \cup A_2 \cup A_3 \cup A_4) \\
 &= P[(B_1 \cap A) \cup (B_2 \cap A) \cup (B_3 \cap A) \cup (B_4 \cap A)] \\
 &= P(B_1 \cap A) + P(B_2 \cap A) + P(B_3 \cap A) + P(B_4 \cap A) \text{ by the Basic Conditioning Rule yields} \\
 &= P(B_1)P(A|B_1) + P(B_2)P(A|B_2) + P(B_3)P(A|B_3) + P(B_4)P(A|B_4) \\
 &= \sum_{j=1}^4 P(B_j)P(A|B_j)
 \end{aligned}$$

The generalization of the above discussion yields The Law of Total Probability.

Theorem: The Law of Total Probability 7.1.2

Theorem: Suppose the events $B_1, B_2, \dots, B_k$ forms a partition on a sample space S such that $P(B_j) > 0$ for $j = 1, 2, 3, \dots, k$. Then for every event A in S ,

$$P(A) = \sum_{j=1}^k P(B_j)P(A|B_j) \quad (7.1.1)$$

Proof

Since $B_1, B_2, \dots, B_k$ forms a partition on the sample space S , then the events $(B_1 \cap A), (B_2 \cap A), \dots$ and $(B_k \cap A)$ forms a partition on A . That is, $(B_1 \cap A) \cup (B_2 \cap A) \cup \dots \cup (B_k \cap A) = A$ and the events $(B_1 \cap A), (B_2 \cap A), \dots$ and $(B_k \cap A)$ are mutually exclusive. Hence, we obtain the following:

$$\begin{aligned}
 P(A) &= P[(B_1 \cap A) \cup (B_2 \cap A) \cup \dots \cup (B_k \cap A)] \\
 &= P(B_1 \cap A) + P(B_2 \cap A) + \dots + P(B_k \cap A) \text{ by the Basic Conditioning Rule yields} \\
 &= P(B_1)P(A|B_1) + P(B_2)P(A|B_2) + \dots + P(B_k)P(A|B_k) \\
 &= \sum_{j=1}^k P(B_j)P(A|B_j)
 \end{aligned}$$

✓ Example 2 7.1.3

Suppose Player 1 draws a card at random, without replacement, from a regular deck of 52 cards and the card remains facedown. Player 2 then draws a card at random from the deck of 51 cards. Show that the probability that Player 2 draws an ace is $\frac{4}{52}$.

Answer 1

Although this question has a simple counting solution (can you find it?), we will opt for a solution rooted in conditional probabilities. The difficulty in this problem is that we are aware of the card that Player 1 has drawn. That is, either Player 1 drew an ace or Player 1 drew a non-ace. Let

$$B_1 = \{\text{Player 1 drew an ace}\}$$

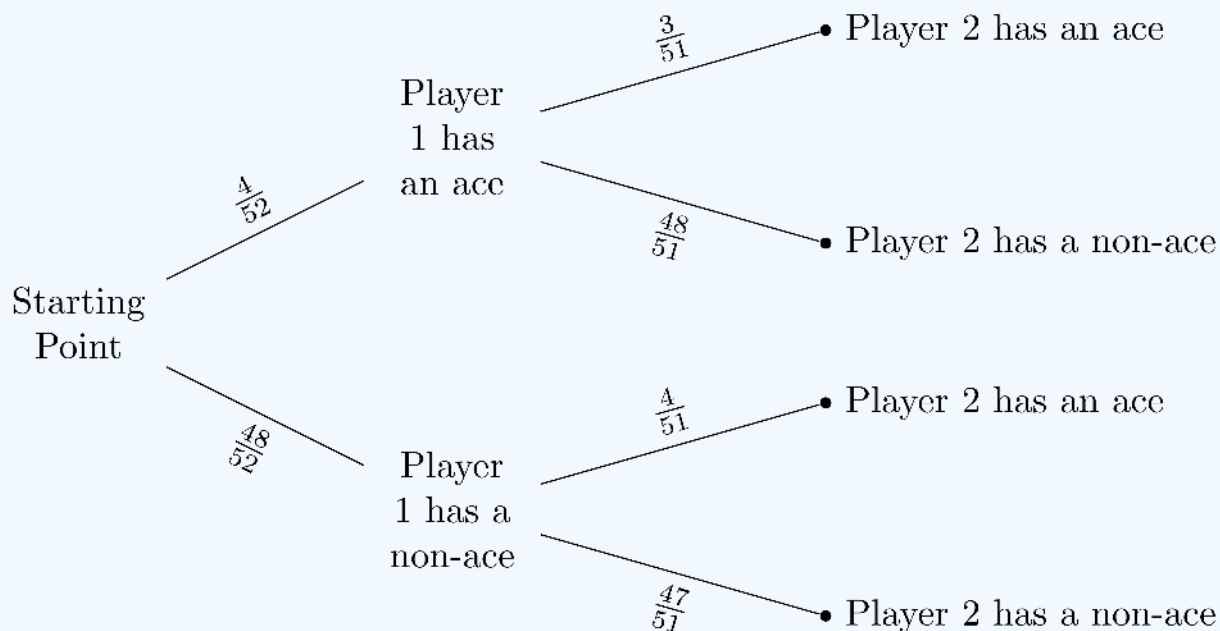
$$B_2 = \{\text{Player 1 drew a non-ace}\}$$

Notice that B_1 and B_2 forms a partition on S . We wish to find the probability that Player 2 draws an ace and so we let $A = \{\text{Player 2 draws an ace}\}$. By The Law of Total Probability, we obtain the following:

$$\begin{aligned} P(A) &= P(B_1)P(A|B_1) + P(B_2)P(A|B_2) \\ &= \frac{4}{52} \left(\frac{3}{51} \right) + \frac{48}{52} \left(\frac{4}{51} \right) \\ &= \frac{4}{52} \\ &= \frac{1}{13} \end{aligned}$$

Answer 2

Alternatively, we may make a tree diagram. To do so, our first level of branches represents our partition while the second level of branches represents the probability of the event we are interested in studying.



We are asked to find the probability that Player 2 has an ace and so we sum the products of the appropriate branches together. Doing so yields:

$$\frac{4}{52} \left(\frac{3}{51} \right) + \frac{48}{52} \left(\frac{4}{51} \right) = \frac{4}{52} = \frac{1}{13}$$

✓ Example 3 7.1.4

Suppose a 15 minute rapid antigen test for the SARS-CoV-2 virus is 80.2% effective in detecting the virus when it is present. However, the test also yields a false positive 8% of the time, meaning if you do not have the disease, then the test will say you are positive for the disease with probability 8%. Assume that 3.39% of people living in Queens, New York has the virus. Suppose a person living in Queens takes the test. Find the probability that the test comes back positive.

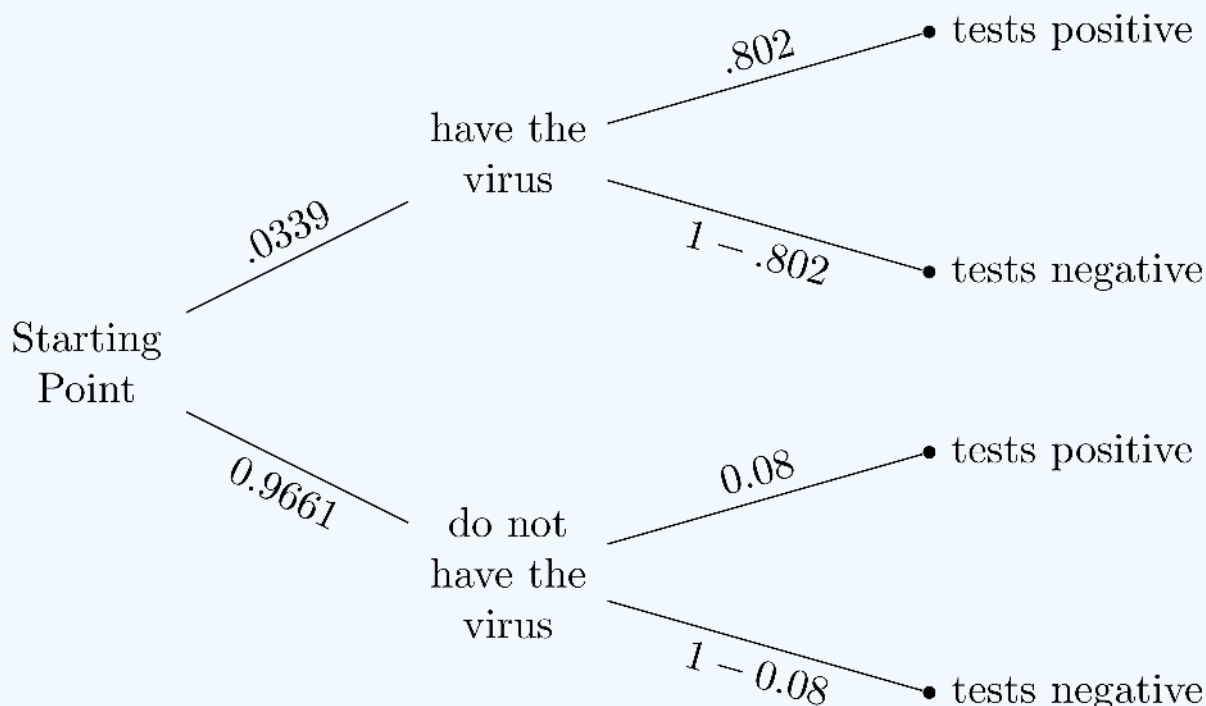
Answer 1

As usual, we first must introduce a partition. Let $B_1 = \{\text{the person has the disease}\}$ and $B_2 = \{\text{the person does not has the disease}\}$. Notice that B_1 and B_2 forms a partition on S . We wish to find the probability that the person has the disease and so let $A = \{\text{the person has the disease}\}$. By The Law of Total Probability, we obtain the following:

$$\begin{aligned}
 P(A) &= P(B_1)P(A|B_1) + P(B_2)P(A|B_2) \\
 &= 0.0339(.802) + 0.9661(0.08) \\
 &\approx 0.1045
 \end{aligned}$$

Answer 2

Alternatively, we may make a tree diagram. To do so, our first level of branches represents our partition while the second level of branches represents the probability of the event we are interested in studying.



We are asked to find the probability that the person tests positive and so we sum the products of the appropriate branches together. Doing so yields:

$$0.0339(.802) + 0.9661(0.08) \approx 0.1045$$

✓ Example 4 7.1.5

A helicopter is missing and it is presumed that the helicopter was equally likely to have crash landed in any of three possible regions. There is a 60% chance the helicopter will be found upon a search of region 1 when the helicopter, is in fact, in that region. This probability is 80% and 20% for regions 2 and 3 respectively. What is the probability that the helicopter will not be found in region 1?

Note: Some students may wonder as to why there isn't a 100% chance the helicopter will be found upon a search of region 1 when the helicopter, is in fact, in that region. The answer is just because you may know an object is located within a certain region, it is not necessarily guaranteed that you will find it! Think about how many times you may have lost an item in your house or apartment. Just because you know the object is in your house, it does not mean that you are guaranteed to find it! There is some probability that you will overlook the item. Similarly, the same thing is true in real life. Just because we may know that the helicopter crash landed in region 1, it is not guaranteed that we will find the helicopter when we search region 1. This is due to environmental and geographic features. For instance, if region 1 is a heavily forested area, then it may be difficult to spot the helicopter from above.

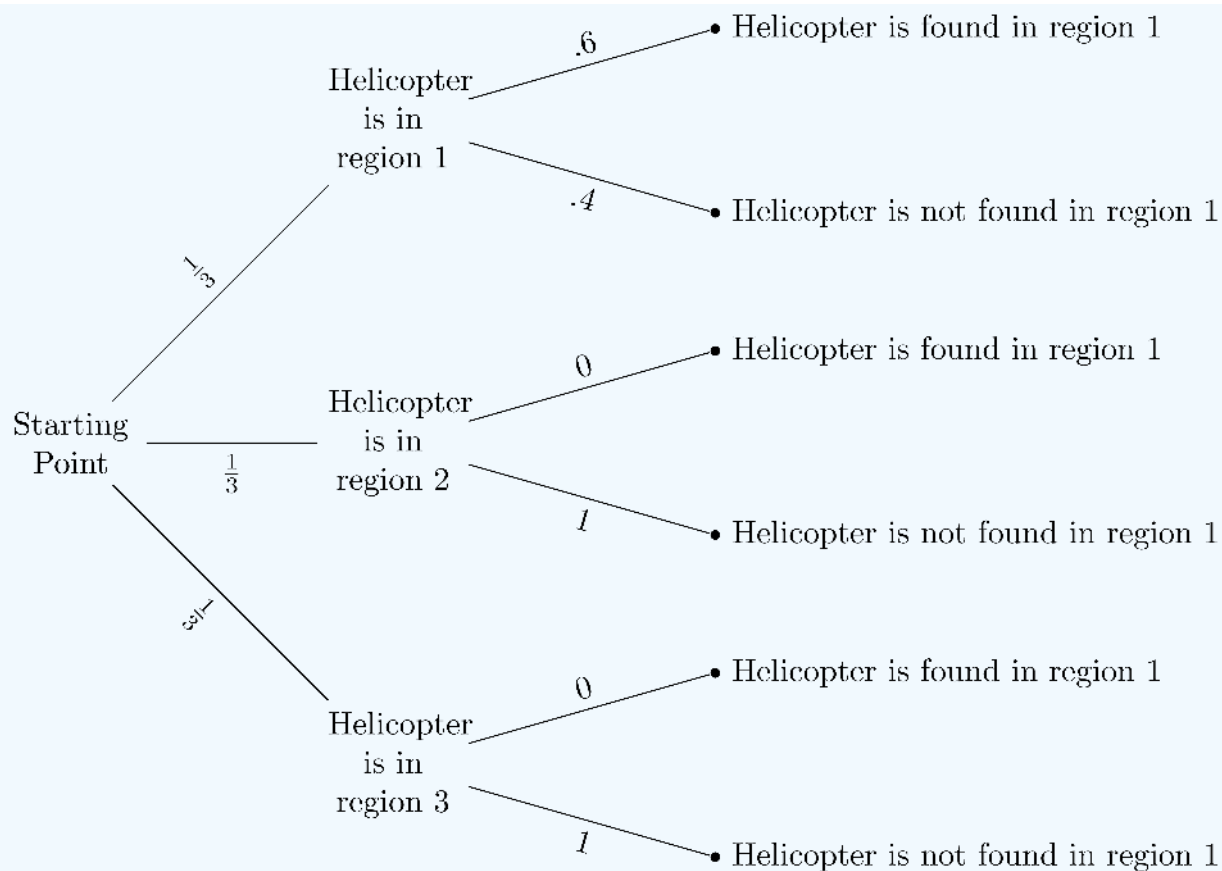
Answer 1

As usual, we must first introduce a partition. Let $B_i = \{\text{the helicopter landed in region } i\}$. Notice that the events B_1, B_2 and B_3 forms a partition on S . We wish to find the probability that the helicopter will not be found in region 1 and so let $A = \{\text{the helicopter is not found in region 1}\}$. By The Law of Total Probability, we obtain the following:

$$\begin{aligned}
 P(A) &= P(B_1)P(A|B_1) + P(B_2)P(A|B_2) + P(B_3)P(A|B_3) \\
 &= \frac{1}{3}(0.4) + \frac{1}{3}(1) + \frac{1}{3}(1) \\
 &= 0.8
 \end{aligned}$$

Answer 2

Alternatively, we may make a tree diagram. To do so, our first level of branches represents our partition while the second level of branches represents the probability of the event we are interested in studying.

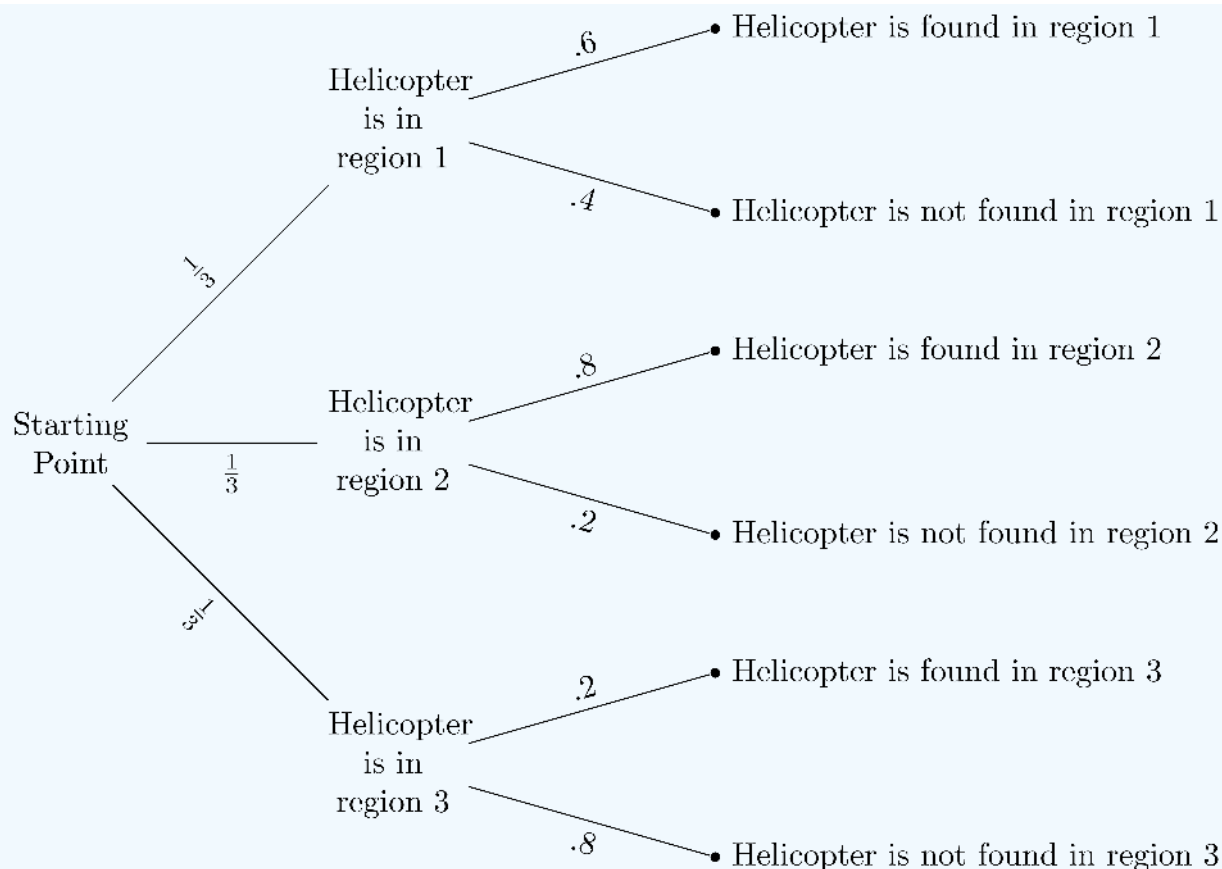


We are asked to find the probability that the helicopter is not found in region 1 and so we sum the products of the appropriate branches together. Doing so yields

$$\frac{1}{3}(0.4) + \frac{1}{3}(1) + \frac{1}{3}(1) = 0.8$$

Answer 3

Alternatively, we may make the probability tree where the first level still represents the partition, but the second level is determined by the information directly stated in the question.



We are asked to find the probability that the helicopter is not found in region 1 and so we sum the products of the appropriate branches together. Doing so yields

$$\frac{1}{3}(0.4) + \frac{1}{3}(0.8) + \frac{1}{3}(0.2) + \frac{1}{3}(0.2) + \frac{1}{3}(0.8) = 0.8$$

Note: You may equivalently note that $1 - \frac{1}{3}(0.6) = 1 - 0.2 = 0.8$.

With The Law of Total Probability under our belts, we are now in a position to state one of the most important results in probability and statistics - Bayes' Theorem.

Theorem: Bayes' Theorem 7.1.6

Theorem: Suppose the events $B_1, B_2, \dots, B_k$ forms a partition on a sample space S such that $P(B_j) > 0$ for $j = 1, 2, \dots, k$ and also suppose that A is an event in S such that $P(A) > 0$. Then for $i = 1, 2, \dots, k$,

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{\sum_{j=1}^k P(B_j)P(A|B_j)} \quad (7.1.2)$$

Proof:

$$\begin{aligned} P(B_i|A) &= \frac{P(B_i \cap A)}{P(A)} \\ &= \frac{P(B_i)P(A|B_i)}{P(A)} \\ &= \frac{P(B_i)P(A|B_i)}{\sum_{j=1}^k P(B_j)P(A|B_j)} \end{aligned}$$

✓ Example 5 7.1.7

Suppose a 15 minute rapid antigen test for the SARS-CoV-2 virus is 80.2% effective in detecting the virus when it is present. However, the test also yields a false positive 8% of the time. Assume that 3.39% of people living in Queens, New York has the virus. Suppose a person living in Queens takes the test and it is learned that the test comes back positive. Find the probability that the person actually has the disease.

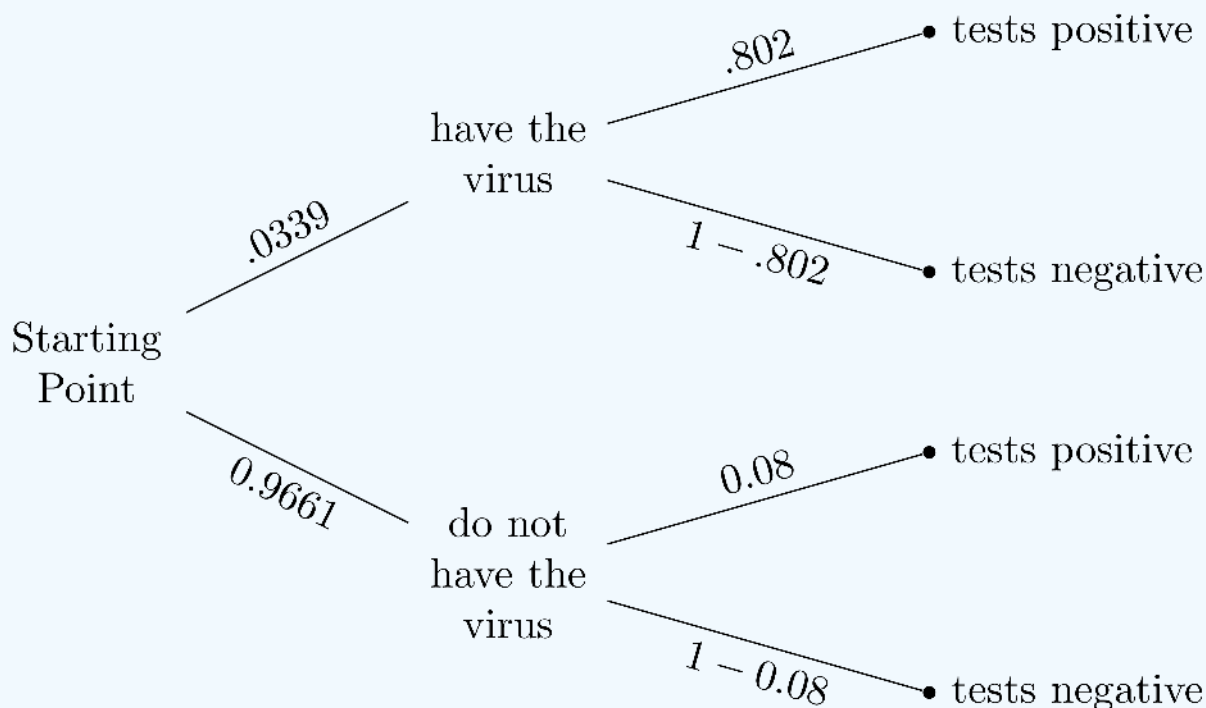
Answer 1

We will stick to the same notation used in Example 3. Let $B_1 = \{\text{the person has the disease}\}$, $B_2 = \{\text{the person does not has the disease}\}$ and $A = \{\text{the person has the disease}\}$. We are asked to find $P(B_1|A)$. By Bayes' Theorem,

$$\begin{aligned} P(B_1|A) &= \frac{P(B_1)P(A|B_1)}{\sum_{j=1}^2 P(B_j)P(A|B_j)} \\ &= \frac{P(B_1)P(A|B_1)}{P(B_1)P(A|B_1) + P(B_2)P(A|B_2)} \\ &= \frac{0.0339(.802)}{0.0339(.802) + 0.9661(0.08)} \\ &\approx 0.2602 \end{aligned}$$

Answer 2

Alternatively, we may use our diagram to answer this question.



We are asked to find $P(\text{have the virus}|\text{tested positive})$. By Bayes' Theorem, we obtain the following:

$$\begin{aligned} P(\text{have the virus}|\text{tested positive}) &= \frac{P(\text{has the virus AND tests positive})}{P(\text{tests positive})} \\ &= \frac{0.0339(.802)}{0.0339(.802) + 0.9661(0.08)} \\ &\approx 0.2602 \end{aligned}$$

✓ Example 7.1.8

A helicopter is missing and it is presumed that the helicopter was equally likely to have crash landed in any of three possible regions. There is a 60% chance the helicopter will be found upon a search of region 1 when the helicopter, is in fact, in that region. This probability is 80% and 20% for regions 2 and 3 respectively. What is the probability that the helicopter is in region 1 given that the plane was not found upon an initial search of region 1?

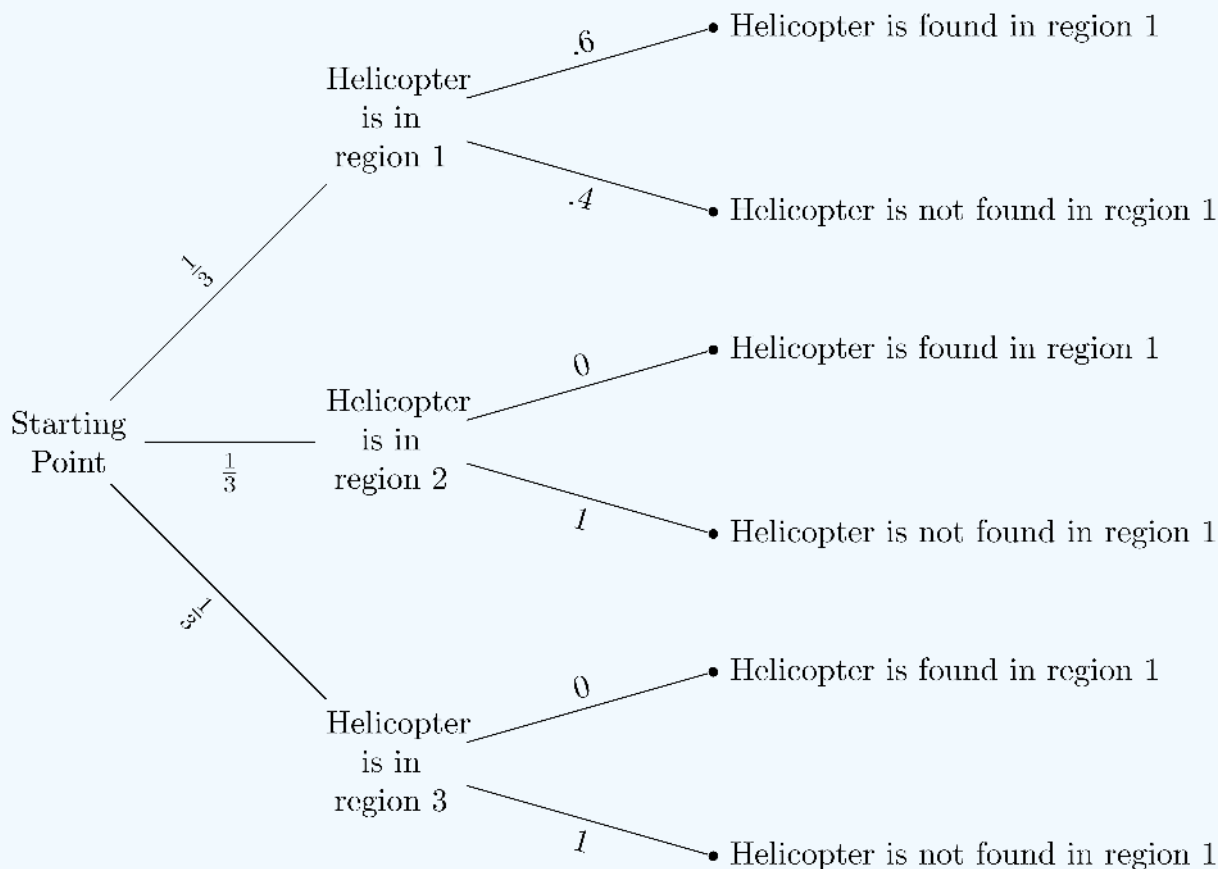
Answer 1

We stick to the same notation as in Example 4. Let $B_i = \{\text{the helicopter landed in region } i\}$ and let $A = \{\text{the helicopter is not found in region 1}\}$. By Bayes' Theorem, we obtain the following:

$$\begin{aligned}
 P(B_1|A) &= \frac{P(B_1 \cap A)}{P(A)} \\
 &= \frac{P(B_1 \cap A)}{P(B_1)P(A|B_1) + P(B_2)P(A|B_2) + P(B_3)P(A|B_3)} \\
 &= \frac{\frac{1}{3} \times 0.4}{\left(\frac{1}{3} \times 0.4\right) + \left(\frac{1}{3} \times 1\right) + \left(\frac{1}{3} \times 1\right)} \\
 &= \frac{1}{6}
 \end{aligned}$$

Answer 2

Alternatively, we may use the following tree diagram.

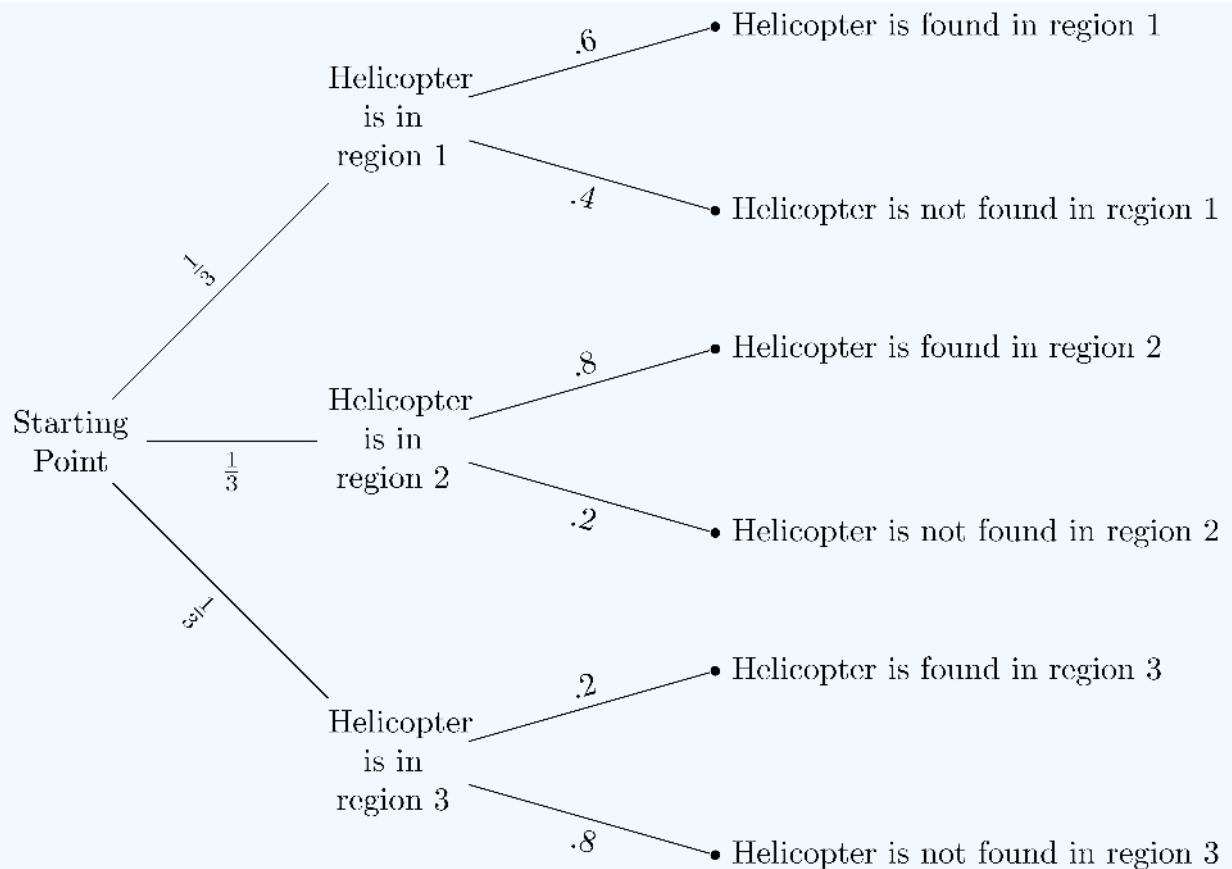


We are asked to find $P(\text{the helicopter is in region 1} | \text{the helicopter is not found in region 1})$. By Bayes' Theorem, we obtain the following:

$$\begin{aligned}
 P(\text{the helicopter is in region 1} | \text{the helicopter is not found in region 1}) &= \frac{P(\text{helicopter is in region 1 AND not the helicopter is not found in region 1})}{P(\text{the helicopter is not found in region 1})} \\
 &= \frac{\frac{1}{3} \times 0.4}{\frac{1}{3}(0.4) + \frac{1}{3}(1) + \frac{1}{3}(1)} \\
 &= \frac{1}{6}
 \end{aligned}$$

Answer 3

With regards to the alternative tree diagram:



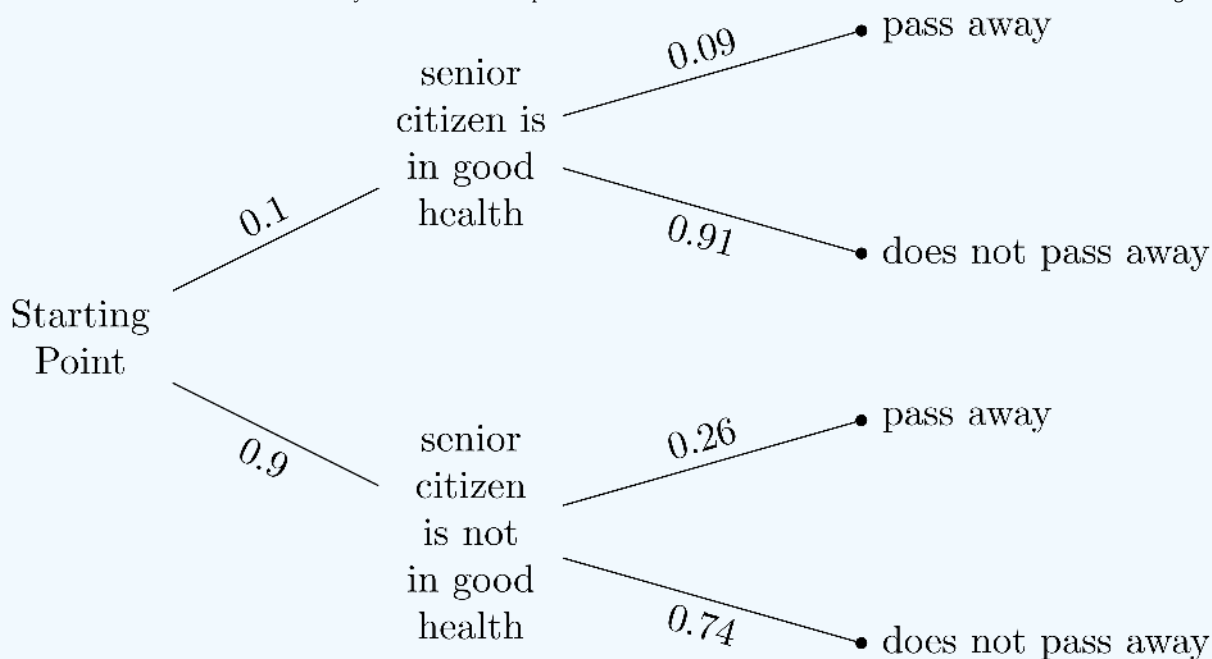
$$\begin{aligned}
 P(\text{the helicopter is in region 1} | \text{the helicopter is not found in region 1}) &= \frac{P(\text{helicopter is in region 1 AND not the helicopter is not found in region 1})}{P(\text{the helicopter is not found in region 1})} \\
 &= \frac{\frac{1}{3} \times 0.4}{\frac{1}{3}(0.4) + \frac{1}{3}(0.8) + \frac{1}{3}(0.2) + \frac{1}{3}(0.2) + \frac{1}{3}(0.8)} \\
 &= \frac{1}{6}
 \end{aligned}$$

✓ Example 7.1.9

A life insurance agency believes that its clients who are senior citizens can be divided into two classes: those who are in good health and those who are not in good health. The agency reports that a senior citizen in good health will pass away within a one year period with probability 0.09 whereas this probability is 0.26 for the senior citizen who is not in good health. Suppose that 10% of the agency's senior citizens are in good health. Given that a senior citizen passed away, what is the probability that the senior citizen was not in good health?

Answer

Here, we only present the tree diagram.



Looking at the above diagram, we see that

$$\begin{aligned}
 P(\text{senior citizen is not in good health} | \text{senior citizen passed away}) &= \frac{P(\text{senior citizen is not in good health AND passes away})}{P(\text{senior citizen passed away})} \\
 &= \frac{0.9(0.26)}{0.1(0.09) + (0.9)(0.26)} \\
 &\approx 0.9630
 \end{aligned}$$

We will end this section with a discussion of The Monty Hall Problem.

✓ Example: The Monty Hall Problem 7.1.10

Suppose that you are a contestant on a game show and you have to choose between three closed doors. Whatever is behind the chosen door is your prize. Behind one door is a brand new car which is what you would like to have and behind the other two doors are goats. The host of the show, Monty, knows where the cars and goats are located, but you, being the contestant, does not. Suppose you pick a door at random. For concreteness, let us suppose that you pick Door 1. Before Monty opens the door you chose, he randomly opens up a different door which reveals a goat behind it. (Note that he never opens the door with the car first). For concreteness, let us suppose that Monty opens up Door 3 to reveal the goat. At this point in time, there are only two doors remaining - your door, Door 1, and the other door, Door 2. Monty then asks you if you would like to keep your door or if you would like to switch to the remaining door. What should you do and why?

Take a moment and think about the question before reading further!

⚠ Caution

Students who are unfamiliar with the above problem often have the following thought process: *after Monty reveals the door there are only two possible outcomes:*

1. My initial door has the car and the other door has the goat.
2. My initial door has the goat and the other door has the car.

Hence if I keep my door, there is a 1 out of 2 chance or a 50% chance that I win. Similarly, if I switch, there is a 50% chance that I win and so I am indifferent to switching since everything is equally likely.

If your thought process was similar to the above argument, then you are not alone.

According to Wikipedia, (https://en.Wikipedia.org/wiki/Monty_Hall_problem):

"The **Monty Hall problem** is a brain teaser, in the form of a probability puzzle, loosely based on the American television game show *Let's Make a Deal* and named after its original host, Monty Hall. The problem was originally posed (and solved) in a letter by Steve Selvin to the *American Statistician* in 1975. It became famous as a question from reader Craig F. Whitaker's letter quoted in Marilyn vos Savant's "Ask Marilyn" column in *Parade* magazine in 1990:

Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say

No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice?

Savant's response was that the contestant should switch to the other door. Under the standard assumptions, the switching strategy has a $2/3$ probability of winning the car, while the strategy of sticking with the initial choice has only a $1/3$ probability.

Many readers of Savant's column refused to believe switching is beneficial and rejected her explanation. After the problem appeared in *Parade*, approximately 10,000 readers, including nearly 1,000 with PhDs, wrote to the magazine, most of them calling Savant wrong. Even when given explanations, simulations, and formal mathematical proofs, many people still did not accept that switching is the best strategy. [Paul Erdős](#), one of the most prolific mathematicians in history, remained unconvinced until he was shown a computer simulation demonstrating Savant's predicted result.

The problem is a paradox of the *veridical* type, because the solution is so counterintuitive it can seem absurd but is nevertheless demonstrably true."

Let us suppose that there is a 50% chance of winning if we stay with our door. This means if we were to conduct the experiment 1000 times, we would expect to win, on average, 500 of time. Let us see if that is indeed true by running a simulation in Java.

Using a Java compiler such as <https://www.jdoodle.com/online-java-compiler/>, run the following code provided by Miguel R. Osorio, Queens College:

Code

```
import java.util.*;

public class Main
{
    public static void main(String args[])
    {
        //this variable holds the number of correct outcomes from running our simulations
        int correctOutcomes = 0;

        //this is the amount of simulations we want to run
        int numSimulations = 1000;

        //this for loop runs our simulation of a single monty hall problem, then, if our simulation
        //was successful and we increase our variable for the amount of correct outcomes by 1 otherwise
        //we do not add anything
        //we do this process "numSimulations" times
        for(int i = 0; i < numSimulations; i++)
        {
            if(runSimOnce() == 1)
                correctOutcomes++;

            //correctOutcomes += (runSimOnceWithSwap() == 1) ? 1 : 0; this is another way to do this but
            //it is less efficient
        }

        System.out.println("Number of times we win the car if we stay with our door (out of 1000): " + correctOutcomes);
        System.out.println("Number of times we lose the car if we stay with our door (out of 1000): " + (numSimulations - correctOutcomes));

        //reset the number of correct outcomes after running the simulation without swapping
        correctOutcomes = 0;

        //this for loop runs our simulation of a single monty hall problem, then, if our simulation
        //was successful and we increase our variable for the amount of correct outcomes by 1 otherwise
        //we do not add anything
        //we do this process "numSimulations" times
        for(int i = 0; i < numSimulations; i++)
        {
            if(runSimOnceWithSwap() == 1)
                correctOutcomes++;

            //correctOutcomes += (runSimOnceWithSwap() == 1) ? 1 : 0; this is another way to do this but
            //it is less efficient
        }

        System.out.println("\nNumber of times we win the car if we switch our door (out of 1000): " + correctOutcomes);
```

```
System.out.println("Number of times we lose the car if we switch our door (out of 1000): " + (numSimulations - correctOutcomes));
System.out.println("Total number of simulations ran: " + (numSimulations * 2));
System.out.println(numSimulations + " ran with staying aka NOT switching");
System.out.println(numSimulations + " ran with switching");
}
/**
 * @function sunRimOnce
 *
 * this function simulates the monty hall problem by choosing randomly, and without revealing
 * one of our wrong choices, giving around a 33% chance of success and returning the outcomes as
 * an integer where 1 means we successfully picked the gate that had the car and 0 means we picked
 * wrong.
 */
public static int runSimOnce()
{
    //this variable is our set of all gates, all of them start at 0 meaning the car is not behind
    //that door
    int[] gates = {0, 0, 0};
    //calls our random number generator so that we can use it
    Random r = new Random();
    /*this chooses a random gate x where x = random number between 0 and
    * the amount of gates not including the last number so [0,3)
    * out of all the possible gates then assigns it the value 1 or effectively
    * "places the car behind that door"
    */
    gates[r.nextInt(gates.length)] = 1;
    /*
    * this selects a gate at random, all with equal probability regardless of how many simulations
    * are run
    */
    int choice = r.nextInt(gates.length);
    /*
    * return 1 or "success" if the content or value of our "gate" is the car(1) or 0 otherwise
    */
    return (gates[choice] == 1) ? 1 : 0;
}
/**
 * @function sunRimOnceWithSwap
 *
 * this function simulates the monty hall problem by choosing randomly, and then swapping
 * our original choice after one of the incorrect gates was revealed.
 *
 * @return 1 if we got the car or 0 if we did not get it after swapping
 */
public static int runSimOnceWithSwap()
{
    //this variable is our set of all gates, all of them start at 0 meaning the car is not behind
    //that door
    int[] gates = {0, 0, 0};
```

```
//calls our random number generator so that we can use it
Random r = new Random();

/*this chooses a random gate x where x = random number between 0 and
* the amount of gates not including the last number so [0,3)
* out of all the possible gates then assigns it the value 1 or effectively
* "places the car behind that door"
*/
gates[r.nextInt(gates.length)] = 1;

/*
* this selects a gate at random, all with equal probability regardless of how many simulations
* are run
*/
int choice = r.nextInt(gates.length);

/*
* this section of the code will first choose which gate to reveal, and change our
* "choice" to the gate that was NOT revealed AND we didn't choose first
*/
if((findReveal(gates, choice) == 2 && choice == 1) || (findReveal(gates, choice) == 1 && choice == 2))
choice = 0;
else if((findReveal(gates, choice) == 2 && choice == 0) || (findReveal(gates, choice) == 0 && choice == 2))
choice = 1;
else if((findReveal(gates, choice) == 1 && choice == 0) || (findReveal(gates, choice) == 0 && choice == 1))
choice = 2;

/*
* return 1 or "success" if the content or value of our "gate"
* after swapping our choice is the car(1) or 0 otherwise
*/
return (gates[choice] == 1) ? 1 : 0;
}

/**
* @function findReveal
* this function finds the gate that we should reveal and returns it as a value bound by [0, length of our set)
*
* @param gates this is the set of total gates that we first started with
* @param originalChoice this is the index of the "gate" that we originally chose
* @return the gate that is NOT the one we chose AND does not have the car behind it
*/
public static int findReveal(int[] gates, int originalChoice)
{
int reveal = 0;
for(int i = 0; i < gates.length; i++)
reveal = (gates[i] == 0 && i != originalChoice) ? i : reveal;
return reveal;
}
}
```

We see that if we were to stay with our door, we only win a little over 300 times. However, if we were to switch with our door, we win will over 600 times. How can we see this mathematically?

✓ Example 7.1.11

Suppose that you are a contestant on a game show and you have to choose between three closed doors. Whatever is behind the chosen door is your prize. Behind one door is a brand new car which is what you would like to have and behind the other two doors are goats. The host of the show, Monty, knows where the cars and goats are located, but you, being the contestant, does not. Suppose you pick a door at random. For concreteness, let us suppose that you pick Door 1. Before Monty opens the door you chose, he randomly opens up a different door which reveals a goat behind it. (Note that he never opens the door with the car first). For concreteness, let us suppose that Monty opens up Door 3 to reveal the goat. At this point in time, there are only two doors remaining - your door, Door 1, and the other door, Door 2. Monty then asks you if you would like to keep your door or if you would like to switch to the remaining door. What should you do and why?

Answer

As usual, we introduce a partition. Let $C_i = \{\text{the car is behind Door } i\}$ and let $A = \text{Monty opened up Door 3}$. We wish to find the probability that the car is behind Door 1 given this information A and compare it to the probability that the car is behind Door 2 given this information A . By Bayes Theorem, we obtain the following:

$$\begin{aligned} P(C_1|A) &= \frac{P(C_1)P(A|C_1)}{P(C_1)P(A|C_1) + P(C_2)P(A|C_2) + P(C_3)P(A|C_3)} \\ &= \frac{\frac{1}{3}P(A|C_1)}{\frac{1}{3}P(A|C_1) + \frac{1}{3}P(A|C_2) + \frac{1}{3}P(A|C_3)} \\ &= \frac{\frac{1}{3}\left(\frac{1}{2}\right)}{\frac{1}{3}\left(\frac{1}{2}\right) + \frac{1}{3}(1) + \frac{1}{3}(0)} \\ &= \frac{1}{3} \end{aligned}$$

This shows us that the probability that the car is behind Door 1, given that Monty opened up Door 3 is $\frac{1}{3}$. Thus the probability that the car is behind Door 2, given that Monty opened up Door 3 must be $\frac{2}{3}$. Alternatively, note by Bayes' Theorem that

$$\begin{aligned} P(C_2|A) &= \frac{P(C_2)P(A|C_2)}{P(C_1)P(A|C_1) + P(C_2)P(A|C_2) + P(C_3)P(A|C_3)} \\ &= \frac{\frac{1}{3}P(A|C_2)}{\frac{1}{3}P(A|C_1) + \frac{1}{3}P(A|C_2) + \frac{1}{3}P(A|C_3)} \\ &= \frac{\frac{1}{3}(1)}{\frac{1}{3}\left(\frac{1}{2}\right) + \frac{1}{3}(1) + \frac{1}{3}(0)} \\ &= \frac{2}{3} \end{aligned}$$

Hence, we are twice as likely to win the car by switching doors.

If you are still not convinced of the result proven above, allow us to make a simple observation about the Monty Hall problem. We can have one of two strategies:

Strategy 1: Our strategy is to switch doors.

Strategy 2: Our strategy is to keep our door.

Under Strategy 1, notice that we *always win by switching* is if we select the door with a goat behind it. This forces Monty to open up the other door with the goat and so by switching we must land on the car.

Under Strategy 2, the *only time we win by keeping our door* is if the initial door we select has the car.

Notice under Strategy 1, there is a $\frac{2}{3}$ chance we select the door with the goat behind it whereas under Strategy 2, there is a $\frac{1}{3}$ chance we select the door with the car. Since we are more likely to select the door with the goat, Strategy 1 is far more beneficial for us.

If you wish to see the tree diagram, then here is one of (many!) explanations: <https://www.youtube.com/watch?v=cphYs1bCeDs>

7.1: Bayes' Theorem is shared under a CC BY license and was authored, remixed, and/or curated by LibreTexts.

CHAPTER OVERVIEW

Chapter 8: Week 8

[8.1: Discrete Random Variables](#)

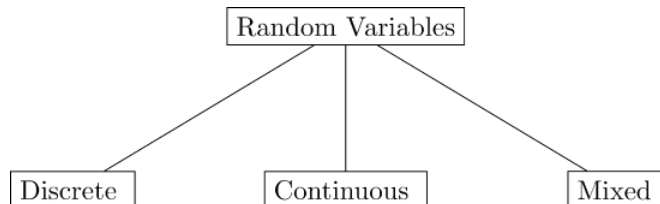
[8.2: The Cumulative Distribution Function](#)

[8.3: Expected Values](#)

Chapter 8: Week 8 is shared under a [CC BY](#) license and was authored, remixed, and/or curated by LibreTexts.

8.1: Discrete Random Variables

As a brief recap of our journey thus far, we first laid down the foundations of probability and then developed a set of theorems which stems from our axiomatic foundation. Afterwards, we related counting techniques to probability and then we related those counting techniques to conditional probabilities. What we are going to do now is make an important jump - we are going to discuss what is meant by a *random variable*. Before we formally define the term, random variable, allow us to make a note. When it comes to random variables, there are three classes of random variables.

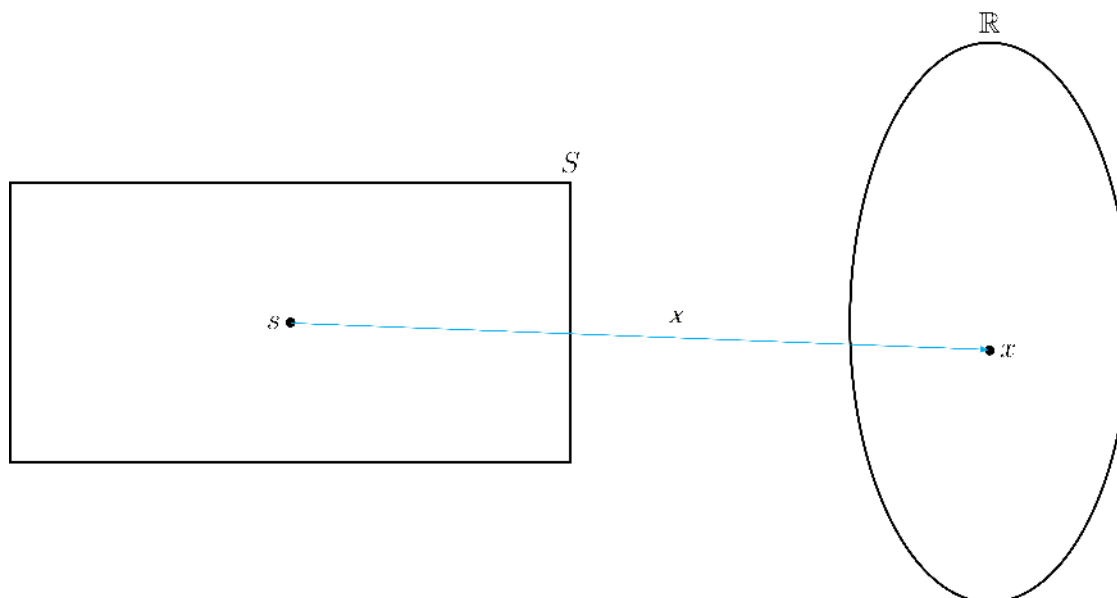


We may have either a discrete random variable, a continuous random variable or a mixed random variable. Once we know how to deal with one branch of random variables, the theory concerning the other two branches are very similar. Hence, we will devote a substantial amount of time discussing discrete random variables and then we will see how the theory for discrete random variables can be generalized to continuous random variables and mixed random variables.

With that said, allow us to formally define what is meant by a random variable.

Definition: Random Variable

Definition: Let $\mathcal{S}$ denote the sample space of an experiment. A random variable is a real-valued function defined on $\mathcal{S}$. Often, we will denote a random variable by a capital letter such as X .



In terms of our usual notation for functions, note that $X: \mathcal{S} \rightarrow \mathbb{R}$. That is, X takes elements in $\mathcal{S}$ and sends them to an element in $\mathbb{R}$. We typically denote an element in $\mathcal{S}$ by s and we will denote an element in $\mathbb{R}$ by x . Hence, $X(s) = x$. Often, x is called a *realization* of X and this will be further explained a moment. In order to illustrate what is going on, allow us to consider a concrete example.

✓ Example $\text{\PageIndex{1}}$

Suppose an experiment consists of flipping a fair coin ten times.

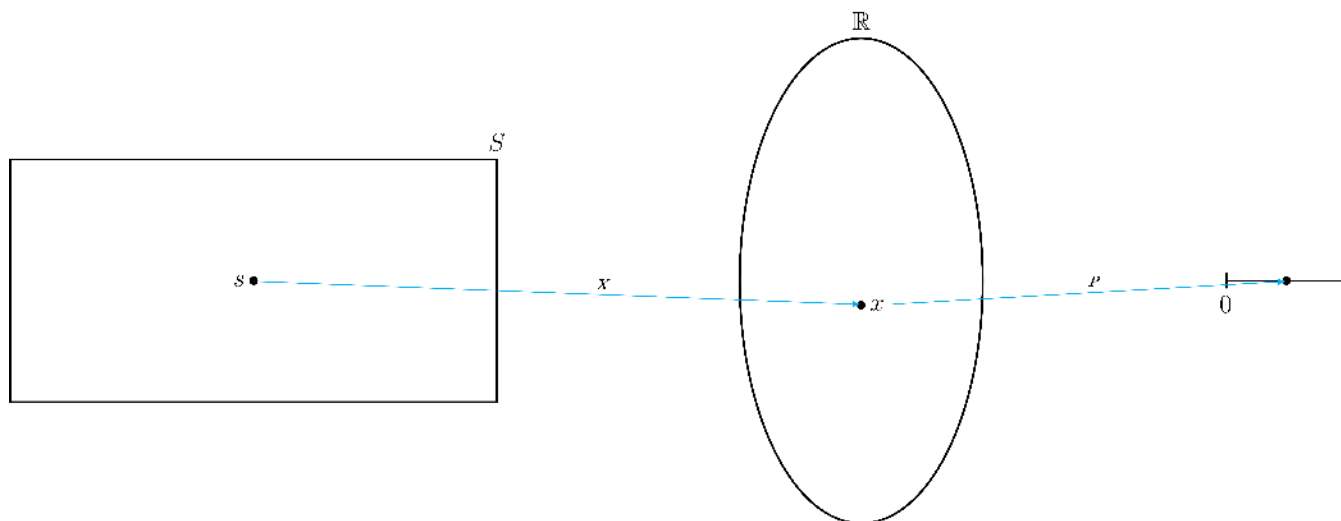
1. What does the sample space for this experiment look like?
2. Write down a few outcomes in $\mathcal{S}$.
3. Allow us to define the X to be the random variable which models the number of heads obtained. Determine the value of X when X is evaluated at the points in $\mathcal{S}$ from Part 2.

4. What are all the possible values of X ?
5. Does X always take the same value?

Answer

1. Let S denote the set of all outcomes where an outcome is an ordered 10-tuple where each entry is either an H for Heads, or a T for tails.
2. Here are some outcomes in S : $(H, H, H, H, H, H, H, H, H, H)$, $(H, T, H, T, H, T, H, T, H, T)$, $(H, H, H, H, H, T, T, T, T, T)$, $(T, T, T, T, T, T, T, T, T, T)$.
3. Observe that $\sum_{i=1}^{10} X_i = 10$ and $\sum_{i=1}^{10} X_i = 5$ and $\sum_{i=1}^{10} X_i = 0$.
4. Since we are flipping the coin ten times, then the number of possible heads are 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10. Hence $X \in \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$.
5. Clearly, as demonstrated above, X does not always take on the same value. It is possible that after flipping the coin ten times, we obtain 0 heads while other times we may obtain 1 head.

We are now going to take an important leap: notice in Question 5, that X does not always take the same value. We see sometimes that $X(s) = 0$, $X(s) = 1$, ..., or $X(s) = 10$. Since the value of X is not always a single, fixed number, we may begin to assign a probability that our random variable, X takes each value. That is, we can find the probability of the event that $X(s) = 0$ and we can find the probability of the event that $X(s) = 1$ and so on.



✓ Example 1

In the example above, find the probability that X takes each value in $\mathbb{R}$.

Answer

The key to answering this question is to re-phrase the question in terms of events and use our previous methods. Doing so yields the following:

$$\begin{aligned}
 P(X(s) = 0) &= P(\text{we obtain 0 heads}) = \frac{|\text{we obtain 0 heads}|}{|S|} = \frac{\binom{10}{0}}{2^{10}} \\
 P(X(s) = 1) &= P(\text{we obtain 1 head}) = \frac{|\text{we obtain 1 head}|}{|S|} = \frac{\binom{10}{1}}{2^{10}} \\
 P(X(s) = 2) &= P(\text{we obtain 2 heads}) = \frac{|\text{we obtain 2 heads}|}{|S|} = \frac{\binom{10}{2}}{2^{10}} \\
 P(X(s) = 3) &= P(\text{we obtain 3 heads}) = \frac{|\text{we obtain 3 heads}|}{|S|} = \frac{\binom{10}{3}}{2^{10}} \\
 P(X(s) = 4) &= P(\text{we obtain 4 heads}) = \frac{|\text{we obtain 4 heads}|}{|S|} = \frac{\binom{10}{4}}{2^{10}} \\
 P(X(s) = 5) &= P(\text{we obtain 5 heads}) = \frac{|\text{we obtain 5 heads}|}{|S|} = \frac{\binom{10}{5}}{2^{10}} \\
 P(X(s) = 6) &= P(\text{we obtain 6 heads}) = \frac{|\text{we obtain 6 heads}|}{|S|} = \frac{\binom{10}{6}}{2^{10}} \\
 P(X(s) = 7) &= P(\text{we obtain 7 heads}) = \frac{|\text{we obtain 7 heads}|}{|S|} = \frac{\binom{10}{7}}{2^{10}} \\
 P(X(s) = 8) &= P(\text{we obtain 8 heads}) = \frac{|\text{we obtain 8 heads}|}{|S|} = \frac{\binom{10}{8}}{2^{10}}
 \end{aligned}$$

$P(X(s) = 9) = P(\text{we obtain 9 heads}) = \frac{\text{we obtain 9 heads}}{|S|} = \frac{\binom{10}{9}}{2^{10}}$.
 $P(X(s) = 10) = P(\text{we obtain 10 heads}) = \frac{\text{we obtain 10 heads}}{|S|} = \frac{\binom{10}{10}}{2^{10}}$.

If we had to summarize the above information, we may compactly write the following:

$$P(X(s) = x) = \frac{\binom{10}{x}}{2^{10}} \sim \text{if } x = 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10$$

Notice, however, we have not answered the question. The above question says to find the probability that (X) takes each possible value in $(\mathbb{R})$. We have only specified the probability (X) takes the values 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10. Seeing that the probability that (X) takes any *other* value is 0, we can further generalize the function to read:

$$P(X(s) = x) = \frac{\binom{10}{x}}{2^{10}} \sim \text{if } x = 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10$$

$$= \begin{cases} \frac{\binom{10}{x}}{2^{10}}, & \text{if } x = 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 \\ 0, & \text{otherwise} \end{cases}$$

Finally, we make one last change to the above function. Recall in calculus, we do not always write $(f(x))$. Instead, we often simply say or write (f) . That is, we do not always have to specify the independent variable since the domain is understood. We adopt the same convention here and instead of writing $(X(s) = x)$, we simply write $(X = x)$.

$$P(X = x) = \frac{\binom{10}{x}}{2^{10}} \sim \text{if } x = 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10$$

$$= \begin{cases} \frac{\binom{10}{x}}{2^{10}}, & \text{if } x = 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 \\ 0, & \text{otherwise} \end{cases}$$

Allow us to interpret the above function. We have written $(P(X(s) = x))$ which means this function returns the probability that our random variable, $(X(s))$ takes a particular value which we denote by (x) . We see that there are only finitely many values for our random variable (X) and this is what is partly meant by a discrete random variable. Moreover, we assign the function $(P(X(s) = x))$ a special name. We call this function the probability mass function of (X) and we denote this function by $(f(x))$. That is,

$$f(x) = P(X = x) = \frac{\binom{10}{x}}{2^{10}} \sim \text{if } x = 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10$$

$$= \begin{cases} \frac{\binom{10}{x}}{2^{10}}, & \text{if } x = 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 \\ 0, & \text{otherwise} \end{cases}$$

And so, allow us to formalize the above terms.

Definition: Discrete Random Variable

Definition: A random variable (X) is said to be a discrete random variable if (X) takes on either a finite number of possible values or a countable infinite number of possible values.

Example $(\text{PageIndex}{1})$

- 1) If $(X: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10)$, then (X) is a discrete random variable since there are a finite number of possibilities for (X) .
- 2) If $(X: 0, 1, 2, 3, \dots)$, then (X) is a discrete random variable since there are a countably infinite number of possible values for (X) .
- 3) If (X) can take any value in the interval $([0,1])$, then (X) is not a discrete random variable since there are an uncountable number of elements in the interval $([0,1])$.

Definition: Probability Mass Function

Definition: For a discrete random variable (X) , we define the probability mass function (p.m.f.) of (X) , often denoted by (f) , to be:

$f(x) = P(X = x)$

Definition: Support/Space of (X)

Definition: For a discrete random variable (X) with probability mass function (f) , we define the support of (X) , or sometimes referred to as the space of (X) , to be $(\{x \in \mathbb{R} : f(x) > 0\})$.

Again, the job of the probability mass function, often denoted by (f) , is tell you the probability that your random variable (X) takes a specific value, (x) . We sometimes refer to (x) as a *realization* of the random variable. The collection of all possible values of (X) that occurs with nonzero probabilities is called the support or space of (X) . For instance, in the above coin toss question, the support of (X) is $(\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\})$.

Note that the graph of the probability mass function looks like the following:

```
success <- 0:10

plot(success,dbinom(success,size=20,prob=.3),
      type='h',
      main='Graph of the PMF',
      ylab='Probability',
      xlab='# of Heads Obtained',
      lwd=3)
options(scipen=999)
dbinom(success, size=20, prob=.3)
```

run restart restart & run all

```
0.000797922662976119 · 0.00683933711122388 · 0.0278458725242687 · 0.0716036722052623 ·
0.130420974373871 · 0.17886305056988 · 0.191638982753443 · 0.164261985217237 · 0.114396739704861 ·
0.0653695655456348 · 0.030817080900085
```

For us, a common question will involve an experiment and a random variable being described. From this description, we will have to find the probability mass function. In order to do so, we apply the following recipe:

Recipe for Deriving a PMF

- 1) Ask yourself, what are the possible values of (X) ? (This gives you the support or space of (X)).
- 2) Find the probabilities that (X) takes these values.
- 3) Summarize these probabilities in a function.

We end this section with a theorem concerning the probability mass function.

Theorem $(\text{PageIndex}\{1\})$

Theorem: Given a discrete random variable (X) , the probability mass function, (f) , must satisfy the following three properties:

- 1) $(f(x) \geq 0)$.
- 2) $(\sum_{\text{all } x \in X(S)} f(x) = 1)$.
- 3) For any event (A) of the sample space (S) , $(\sum_{x \in A} f(x) = P(X \in A))$.

For instance, in the above example with the coin, we see that:

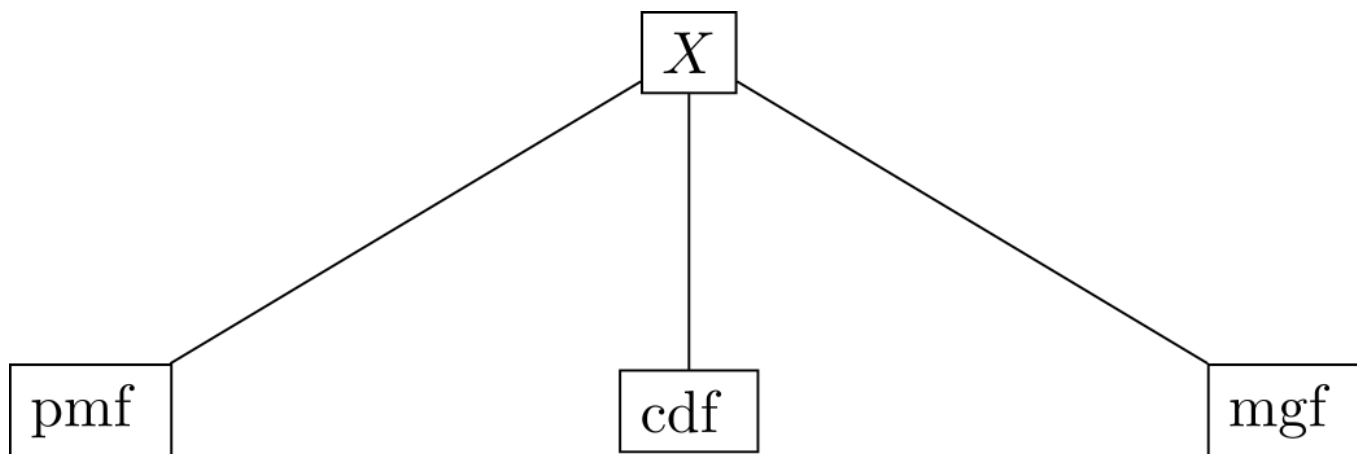
- 1) $(f(x) \geq 0)$ for all (x)
- 2) $(f(0) + f(1) + \dots + f(10) = 1)$
- 3) If $(A = \{\text{we obtained 5 heads or less}\} = \{X \leq 5\})$, then $(P(A) = f(0) + f(1) + f(2) + f(3) + f(4) + f(5))$.

8.2: The Cumulative Distribution Function

To recap the previous section, we defined what is meant by a random variable. A random variable, X , is a function from $S \rightarrow \mathbb{R}$. Afterwards, we then discussed the probability mass function, f , associated to the random variable X .

Note that X , itself, is a function and associated with this function is the probability mass function. In total, there are three important functions associated with a discrete random variable. These functions are

- 1) the probability mass function (pmf)
- 2) the cumulative distribution function (cdf)
- 3) the moment generating function (mgf)



In this section, we will discuss the cumulative distribution function.

Definition: Cumulative Distribution Function

Definition: For a discrete random variable X with probability mass function f , we define the cumulative distribution function (c.d.f.) of X , often denoted by F , to be:

$$F(x) = P(X \leq x), \quad -\infty < x < \infty$$

As a quick comparison, allow us to discuss the difference between a pmf and a cdf. As we recall from the previous section, the job of a pmf is to tell us the probability that X **equals a single, particular** value. Meanwhile, a cdf tells us the probability that X is **less than or equal to a particular value**.

✓ Example 8.2.1

Suppose a random variable X has the following probability mass function.

$$f(x) = P(X = x) = \begin{cases} .855, & \text{if } x = 0 \\ .140, & \text{if } x = 100000 \\ .005, & \text{if } x = 200000 \\ 0, & \text{otherwise} \end{cases}$$

Find the cumulative distribution function of X .

Answer 1

One way to find the cumulative distribution function is to evaluate the cdf at selected points and then generalize our findings. This method can be extremely inefficient, especially if there are many possible values for X . However, this is a good exercise to do at least once since this will give some intuition about how the cumulative distribution function behaves and so we perform this method here. Allow us to find the following:

$F(-100), F(-1), F(-0.01), F(0), F(0.01), F(1), F(50,000), F(99,999.99), F(100,000), F(100,000.01), F(150,000), F(199,999.99), F(200,000), F(1,000,000)$ To do so, we will use the third part of the [theorem](#) from the previous section.

1. $F(-100) = P(X \leq -100) = \sum_{\text{all } x \leq -100} f(x) = 0$
2. $F(-1) = P(X \leq -1) = 0 \sum_{\text{all } x \leq -1} f(x) = 0$
3. $F(-0.01) = P(X \leq -0.01) = \sum_{\text{all } x \leq -0.01} f(x) = 0$
4. $F(0) = P(X \leq 0) = \sum_{\text{all } x \leq 0} f(x) = f(0) = 0.855$

$$5. F(1) = P(X \leq 1) = 0 \sum_{\text{all } x \leq 1} f(x) = f(0) = 0.855$$

$$6. F(50,000) = P(X \leq 50,000) = 0 \sum_{\text{all } x \leq 50,000} f(x) = f(0) = 0.855$$

$$7. F(99,999.99) = P(X \leq 99,999.99) = 0 \sum_{\text{all } x \leq 99,999.99} f(x) = f(0) = 0.855$$

$$8. F(100,000) = P(X \leq 100,000) = 0 \sum_{\text{all } x \leq 100,000} f(x) = f(0) + f(100,000) = 0.855 + 0.140 = 0.995$$

$$9. F(100,000.01) = P(X \leq 100,000.01) = 0 \sum_{\text{all } x \leq 100,000.01} f(x) = f(0) + f(100,000) = 0.855 + 0.140 = 0.995$$

$$10. F(150,000) = P(X \leq 150,000) = 0 \sum_{\text{all } x \leq 150,000} f(x) = f(0) + f(100,000) = 0.855 + 0.140 = 0.995$$

$$11. F(199,999.99) = P(X \leq 199,999.99) = 0 \sum_{\text{all } x \leq 199,999.99} f(x) = f(0) + f(100,000) = 0.855 + 0.140 = 0.995$$

$$12. F(200,000) = P(X \leq 200,000) = 0 \sum_{\text{all } x \leq 200,000} f(x) = f(0) + f(100,000) + f(200,000) = 0.855 + 0.140 + 0.005 = 1$$

$$13. F(1,000,000) = P(X \leq 1,000,000) = 0 \sum_{\text{all } x \leq 1,000,000} f(x) = f(0) + f(100,000) + f(200,000) = 0.855 + 0.140 + 0.005 = 1$$

Based off these computations, we see that for any value $x < 0$, the cdf is zero.

At $x = 0$, the cdf is equal to 0.855.

For any x where $0 \leq x < 100,000$ the cdf should also be equal to 0.855 since the only possible value of X that occurs with a nonzero probability is at $X = 0$.

At $x = 100,000$ the cdf is equal to 0.995.

For any x where $100,000 \leq x < 200,000$ the cdf should also be equal to 0.995 since the only possible values of X that occurs with a nonzero probability is at $X = 0$ and $X = 100,000$.

At $x = 200,000$ the cdf is equal to 1.

For any x where $x \geq 200,000$ the cdf should also be equal to 1 since the possible values of X that occurs with a nonzero probability is at $X = 0$, $X = 100,000$ and $X = 200,000$.

Putting everything together, we see that

$$F(x) = P(X \leq x) = \begin{cases} 0, & \text{if } x < 0 \\ 0.855, & \text{if } 0 \leq x < 100,000 \\ .995, & \text{if } 100,000 \leq x < 200,000 \\ 1, & \text{if } x \geq 200,000 \end{cases}$$

Answer 2

More efficiently, we may use the following remark.

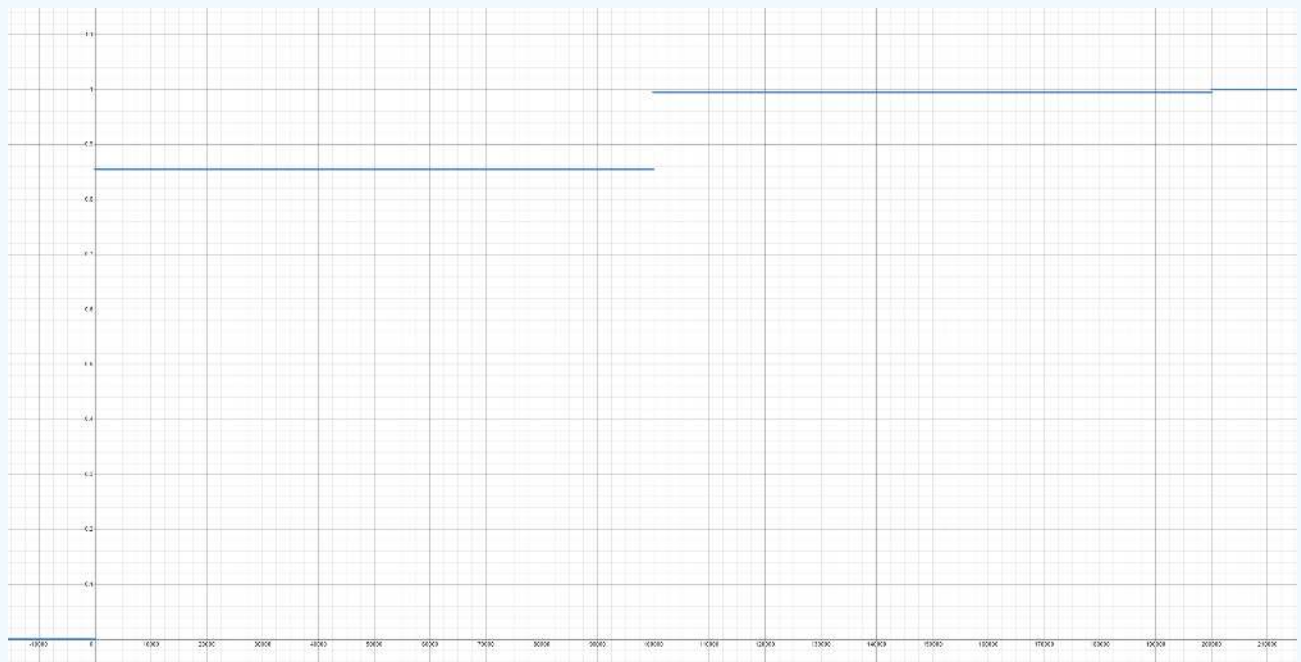
Note

If X is a discrete random variables whose possible values are $x_1, x_2, \dots, x_n$ where $x_1 < x_2 < \dots < x_n$, then the cumulative distribution function of X is a step function that starts off at a "height" of zero, "ends" at a height of 1, and whose jump at x_i is equal to $f(x_i)$.

To use this remark, allow us to graph the pmf. Doing so yields the following image:



Based off our remark, the cumulative distribution function will be a step function that starts at a height of zero, ends at a height of 1, and whose jump in between will occur at each possible value of X . The height of this jump is precisely the height of each red bar.



Looking at the graph, we see that $F(x) = P(X \leq x) = \begin{cases} 0, & \text{if } x < 0 \\ 0.855, & \text{if } 0 \leq x < 100,000 \\ .995, & \text{if } 100,000 \leq x < 200,000 \\ 1, & \text{if } x \geq 200,000 \end{cases}$

8.2: The Cumulative Distribution Function is shared under a CC BY license and was authored, remixed, and/or curated by LibreTexts.

8.3: Expected Values

Now that we are familiar with random variables, we are going to discuss a natural idea - the *average* of a random variable. In fact, we argue that the average of a random variable is something that we all are already familiar with. Consider the following question:

✓ Motivating Example 1 8.3.1

Suppose for a certain college course, the grading policy is the following:

1. Quiz Average - 15%
2. Exam 1 - 20%
3. Exam 2 - 20%
4. Exam 3 - 20%
5. Final Exam - 25%

Let us suppose your quiz average is an 85, your Exam 1 score is 72, your Exam 2 score is 85, your Exam 3 score is 90 and your Final Exam score is 81. How would you compute your average?

Answer

We take every score and multiply the score by the "weight". We then add the resulting products to compute our average for the class. Doing so yields, $\text{Average} = 0.15(85) + 0.2(72) + 0.2(85) + 0.2(90) + 0.25(81) = 82.4$ and so you have obtained a *B*– in the course.

Allow us to now introduce some notation. In the above example, we can denote our score on any assignment as x_i and the corresponding weight outlined in the grading policy can be thought of as $f(x_i)$. Hence, we obtain the following table:

Assignment	x_i	$f(x_i)$	$x_i f(x_i)$
Quiz Average	0.15	85	0.15(85)
Exam 1	0.20	72	0.2(72)
Exam 2	0.20	85	0.2(85)
Exam 3	0.20	90	0.2(90)
Final Exam	0.25	81	0.25(81)

We see that the average is computed by adding up the entries under the column labeled $x_i f(x_i)$. Hence, we see that

$$\text{Average} = \sum_{i=1}^5 x_i f(x_i)$$

And if we were to generalize the above formula, we can conclude that

$$\text{Average} = \sum_{\text{all } x} x f(x) \quad (8.3.1)$$

The above equation is how we precisely define the average of a discrete random variable.

Definition: Expected Value

Definition: For a discrete random variable X with probability mass function f , we define the expected (average) value of X , denoted by $\mathbb{E}[X]$ or by μ , to be

$$\mathbb{E}[X] = \sum_{\text{all } x} x f(x)$$

Allow us to take a look at a few computations involving the expected value of a random variable.

✓ Example 2 8.3.1

Suppose the probability mass function of X is given by $f(x) = P(X = x) = \begin{cases} .855, & \text{if } x = 0 \\ .140, & \text{if } x = 100000 \\ .005, & \text{if } x = 200000 \\ 0, & \text{otherwise} \end{cases}$

Find $\mathbb{E}[X]$.

Answer

To find the average of X , we simply apply the formula, Average = $\sum_{\text{all } x} x f(x)$. That is, we multiply each value of the random variable by its "weight" and add up the products. Doing so yields the following result:

$$\begin{aligned} \mathbb{E}[X] &= \sum_{\text{all } x} x f(x) \\ &= 0(0.855) + 100,000(0.140) + 200,000(0.005) \\ &= 15,000 \end{aligned}$$

✓ Example 3 8.3.1

Suppose X is a discrete random variable whose probability mass function is given by $f(x) = P(X = x) = \begin{cases} cx, & \text{if } x = 1, 2, 3, 4 \\ 0, & \text{otherwise} \end{cases}$

where c denotes some constant.

- 1) Find the value of c .
- 2) Find $\mathbb{E}[X]$.

Answer

1) First note that c cannot be any arbitrary real number. For instance, if we said that $c = 2$ then we are saying that $P(X = 4) = f(4) = 8$ and so we have violated our theorems since we proved a probability cannot be greater than 1.

For questions like these, there is only one, unique value for c . To find the appropriate value of c , we use the [third property](#) of a probability mass function. Since $\sum_{\text{all } x} f(x) = 1$, then we obtain the following:

$$\begin{aligned} \sum_{\text{all } x} f(x) &= 1 \\ \sum_{\text{all } x} cx &= 1 \\ 1c + 2c + 3c + 4c &= 1 \\ 10c &= 1 \\ c &= \frac{1}{10} \end{aligned}$$

And so $c = \frac{1}{10}$.

2) We now know that X has the following probability mass function:

$$f(x) = P(X = x) = \begin{cases} \frac{x}{10}, & \text{if } x = 1, 2, 3, 4 \\ 0, & \text{otherwise} \end{cases}$$

To find $\mathbb{E}[X]$, we simply apply our definition. Doing so yields:

$$\begin{aligned}
 \mathbb{E}[X] &= \sum_{\text{all } x} x f(x) \\
 &= \sum_{x=1}^4 x \frac{x}{10} \\
 &= \sum_{x=1}^4 \frac{x^2}{10} \\
 &= \frac{1^2}{10} + \frac{2^2}{10} + \frac{3^2}{10} + \frac{4^2}{10} \\
 &= \frac{30}{10} \\
 &= 3
 \end{aligned}$$

✓ Example 4 8.3.1

Suppose the probability mass function of X is given by $f(x) = P(X = x) = \begin{cases} \frac{1}{6}, & \text{if } x = 1, 2, 3, 4, 5, 6 \\ 0, & \text{otherwise} \end{cases}$

Find $\mathbb{E}[X]$.

Answer

To find the average of X , we simply apply the formula, $\text{Average} = \sum_{\text{all } x} x f(x)$. That is, we multiply each value of the random variable by its "weight" and add up the products. Doing so yields the following result:

$$\begin{aligned}
 \mathbb{E}[X] &= \sum_{\text{all } x} x f(x) \\
 &= \frac{1}{6}(1) + \frac{1}{6}(2) + \frac{1}{6}(3) + \frac{1}{6}(4) + \frac{1}{6}(5) + \frac{1}{6}(6) \\
 &= \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) \\
 &= 3.5
 \end{aligned}$$

There are two important remarks to make concerning $\mathbb{E}[X]$.

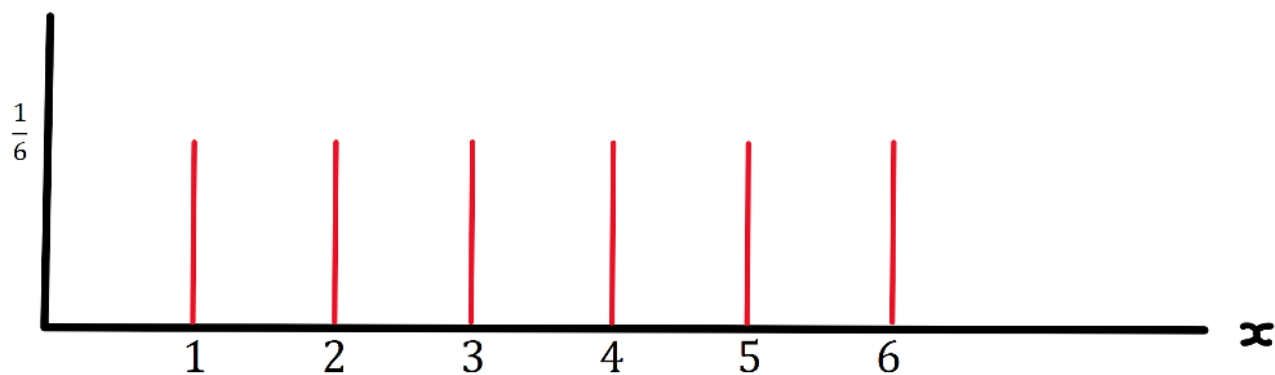
📌 Remarks

1) First note that the average value is not guaranteed to be a possible value of X . In our motivating question, the average was an 82.4, yet we never scored an 82.4 on any of the assignments. The same can be said about the two examples considered above. Notice in Example 2, the average was 15,000 which is not a possible value of X and in Example 3 the average was 3.5 which is again not a possible value of X .

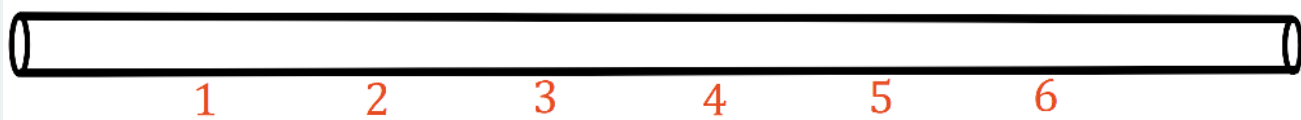
2) From physics, especially classical mechanics, there is a nice way to interpret the expected value. The expected value of a discrete random variable is nothing more than the so called "center of mass" or "balance point". What do we mean by this? \

Suppose X is a discrete random variable with values x_i and corresponding probabilities p_i . Imagine a weightless rod such that at location x_i on the rod, an object of weight p_i is placed. The point at which the rod balances is precisely $\mathbb{E}[X]$.

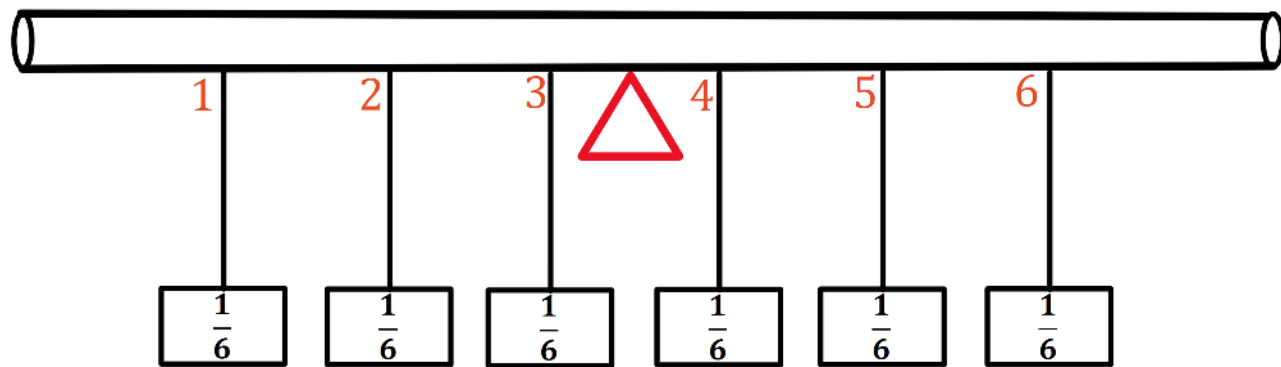
For instance, consider the following bar graph for the last problem.



Imagine the x -axis as a weightless rod and at each 1 unit interval, we attach a weight of $\frac{1}{6}$ pounds.



Where would you have to place your finger to balance this rod on top of? Precisely at the half way mark which is at $\mathbb{E}[X] = 3.5$.



In summary, to find the average of a discrete random variable X , we simply compute $\sum_{\text{all } x} x f(x)$. Note that in order to do so, we require the probability mass function of X . Hence, *at this point, to find the average of a discrete random variable, we need the probability mass function of the random variable.*

The question we will now consider for the remainder of this section is the following: given a random variable X , we know how to compute $\mathbb{E}[X]$. However, what if we wish to compute the average of a function of X ?

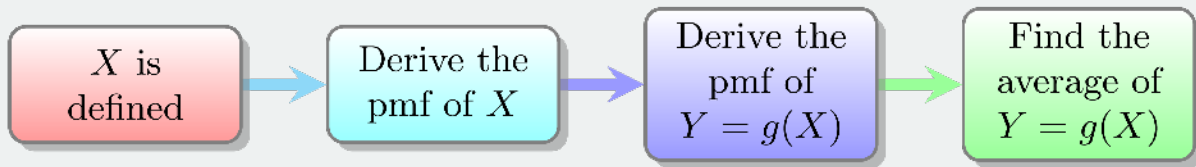
For instance, suppose we defined some random variable X . Instead of computing $\mathbb{E}[X]$, we might be interested in computing $\mathbb{E}[X^2]$ or $\mathbb{E}[5x^3]$.

More formally, the question we are considering is given a random variable X and a real valued function g , how can we find $\mathbb{E}[g(X)]$? Note that $g(X)$ is a new random variable which we will denote by Y and there are two ways in which we can compute the average of $g(X) = Y$.

📌 Method 1 to Compute $\mathbb{E}[u(X)]$.

Recall we stated that *at this point, to find the average of a discrete random variable, we need the probability mass function of the random variable.* Hence, given a random variable X , if we wish to find the average of $g(X) = Y$, it suffices to derive the

probability mass function of Y and use the probability mass function to find the average.



In practice, the above process can be somewhat difficult and so allow us to introduce Method 2 which is an important theorem.

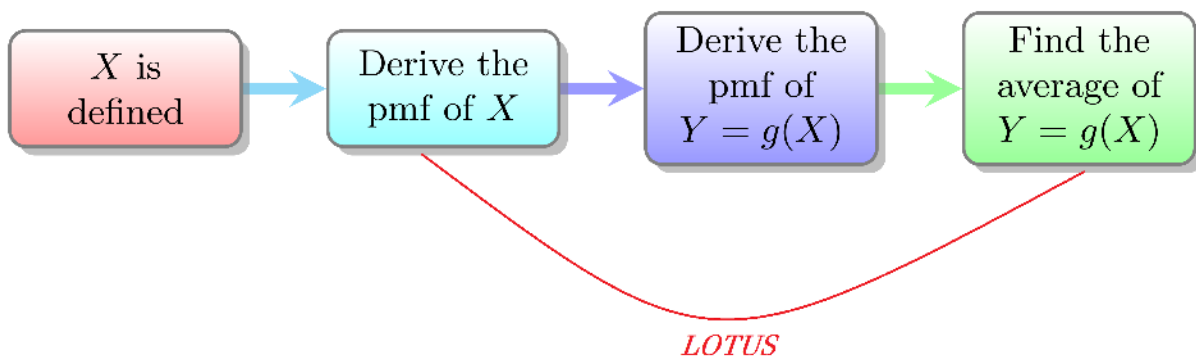
Method 2 to Compute $\mathbb{E}[u(X)]$.

Theorem: The Law of the Unconscious Statistician 8.3.1

Theorem: If X is a discrete random variable with probability mass function f and if $u(X)$ is any real-valued function of X , then

$$\mathbb{E}[u(X)] = \sum_{\text{all } x} u(x)f(x) \quad (8.3.2)$$

We will see that applying The Law of the Unconscious Statistician (The LOTUS), allows us to skip the derivation of the probability mass function of $u(X) = Y$.



Example 5 8.3.1

Suppose X is a discrete random variable whose probability mass function is given by

$$f(x) = P(X = x) = \begin{cases} 0.2, & \text{if } x = -1 \\ 0.5, & \text{if } x = 0 \\ 0.3, & \text{if } x = 1 \\ 0, & \text{otherwise} \end{cases}$$

Find $\mathbb{E}[X^2]$.

Method 1

Looking back at the flowchart presented under Method 1, we see that the random variable X has been defined and the probability mass function of X has already been given. Thus, we now must derive the probability mass function of Y which we will denote by g . Recall how to derive a probability mass function. We first ask ourselves, what are the possible values of Y ? Seeing the $X : -1, 0, 1$ then $Y = X^2 : 1, 0, 1$. So $Y : 0, 1$. We now must find the probabilities that Y takes each value. Note that

$$P(Y = 0) = P(X^2 = 0) = P(X = 0) = f(0) = 0.5 \quad .$$

$$P(Y = 1) = P(X^2 = 1) = P(\{X = -1\} \cup \{X = 1\}) = P(\{X = -1\}) + P(\{X = 1\}) = f(0) + f(1) = 0.2 + 0.3 = 0.5$$

Summarizing everything in a function, we see that $g(y) = P(Y = y) = \begin{cases} 0.5, & \text{if } x = 0 \\ 0.5, & \text{if } x = 1 \\ 0, & \text{otherwise} \end{cases}$

Hence,

$$\mathbb{E}[Y] = \sum_{\text{all } y} yg(y) = 0(0.5) + 1(0.5) = 0.5.$$

Method 2

Alternatively, we may apply The LOTUS. Recall The LOTUS states

$$\mathbb{E}[u(X)] = \sum_{\text{all } x} u(x)f(x) \quad (8.3.3)$$

Hence

$$\begin{aligned} \mathbb{E}[X^2] &= \sum_{\text{all } x} x^2 f(x) \\ &= (-1)^2 f(-1) + 0^2 f(0) + 1^2 f(1) \\ &= f(-1) + f(1) \\ &= 0.2 + 0.3 \\ &= 0.5 \end{aligned}$$

Observe that Method 1 could become a complicated process if the original random variable has many values and if $u(X)$ is not injective whereas Method 2 simply relies on the underlying probability mass function of the random variable in the transformation.

✓ Example 6 8.3.1

Suppose X is a discrete random variable whose probability mass function is given by

$$f(x) = P(X = x) = \begin{cases} \frac{x}{10}, & \text{if } x = 1, 2, 3, 4 \\ 0, & \text{otherwise} \end{cases}$$

Find $\mathbb{E}[X^2]$.

Answer

To find $\mathbb{E}[X^2]$, we apply The LOTUS. Doing so yields:

$$\begin{aligned} \mathbb{E}[X^2] &= \sum_{\text{all } x} x^2 f(x) \\ &= \sum_{x=1}^4 x^2 \frac{x}{10} \\ &= \sum_{x=1}^4 \frac{x^3}{10} \\ &= \frac{1^3}{10} + \frac{2^3}{10} + \frac{3^3}{10} + \frac{4^3}{10} \\ &= \frac{100}{10} \\ &= 10. \end{aligned}$$

Theorem 8.3.1

Theorem: Suppose X is a discrete random variable. Then the following hold:

1) If c is a constant, then $\mathbb{E}[c] = c$.

- 2) If c is a constant, and u is a real-valued function of X then $\mathbb{E}[cu(X)] = c\mathbb{E}[u(X)]$.
- 3) If a and b are constants, then $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$.
- 4) If u and v are real-valued functions of X , then $\mathbb{E}[u(X) + v(X)] = \mathbb{E}[u(X)] + \mathbb{E}[v(X)]$.
- 5) If a and b are constants and u and v are real-valued functions of X , then $\mathbb{E}[au(X) + bv(X)] = a\mathbb{E}[u(X)] + b\mathbb{E}[v(X)]$.

Here is an application of the above theorem.

✓ Example 7 8.3.1

Suppose X is a discrete random variable whose probability mass function is given by

$$f(x) = P(X = x) = \begin{cases} \frac{x}{10}, & \text{if } x = 1, 2, 3, 4 \\ 0, & \text{otherwise} \end{cases}$$

Find $\mathbb{E}[X(5 - X)]$.

Answer

Although we can apply The LOTUS to the above problem, allow us to use the preceding theorem. Doing so yields:

$$\begin{aligned} \mathbb{E}[X(5 - X)] &= \mathbb{E}[5X - X^2] \\ &= \mathbb{E}[5X] - \mathbb{E}[X^2] \\ &= 5\mathbb{E}[X] - \mathbb{E}[X^2] \\ &= 5(3) - 10 \\ &= 5. \end{aligned}$$

To conclude this section, we present a definition which will be important in the next lecture.

Definition: The n^{th} Moment

Definition: Let n denote a positive integer and suppose X is a random variable. We refer to $\mathbb{E}[X^n]$ as the n^{th}  **moment** of X .

This means $\mathbb{E}[X]$ is the first moment of X and $\mathbb{E}[X^2]$ is the second moment of X and so on and so forth.

8.3: Expected Values is shared under a [CC BY](https://creativecommons.org/licenses/by/4.0/) license and was authored, remixed, and/or curated by LibreTexts.

CHAPTER OVERVIEW

Chapter 9: Week 9

[9.1: Variances](#)

[9.2: The Uniform Random Variable](#)

[9.3: The Bernoulli and Binomial Random Variables](#)

[9.4: The Geometric and Negative Binomial Random Variables](#)

[9.5: The Poisson Random Variable](#)

Chapter 9: Week 9 is shared under a [CC BY](#) license and was authored, remixed, and/or curated by LibreTexts.

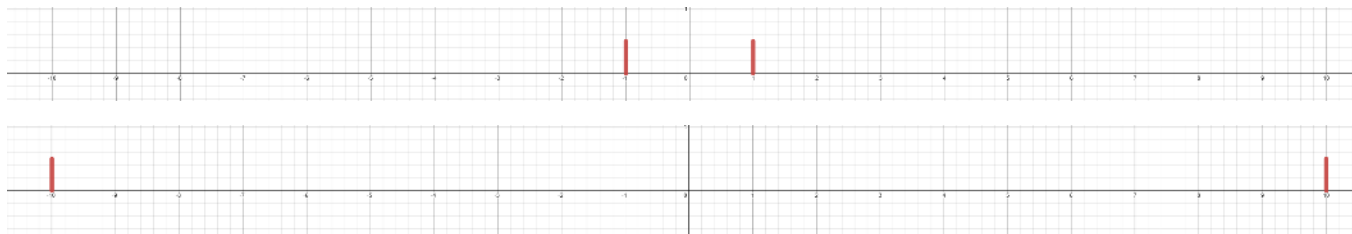
9.1: Variances

The Law of the Unconscious Statistician will play an important role in this section. To motivate the next topic, consider the following two random variables, X and Y with the following probability mass functions:

$$f(x) = P(X = x) = \begin{cases} 0.5, & \text{if } x = -1, 1 \\ 0, & \text{otherwise} \end{cases}$$

$$g(y) = P(Y = y) = \begin{cases} 0.5, & \text{if } y = -100, 100 \\ 0, & \text{otherwise} \end{cases}$$

The graph of the probability mass function looks like the following:



Based off our intuition that $\mathbb{E}[X]$ and $\mathbb{E}[Y]$ represents the center of mass of each distribution, we would perhaps guess that $\mathbb{E}[X] = 0$ and $\mathbb{E}[Y] = 0$. The actual calculation is straightforward and indeed $\mathbb{E}[X] = 0 = \mathbb{E}[Y]$.

Notice that X and Y have the same average, yet there is a striking difference between the two distributions. What is this difference?

The difference is that the distribution of Y is far more spread out than the distribution of X . That is, the values of Y seem to "vary" more than the values of X . Our goal is to develop a measure of how much a distribution can vary. The way we define this measure is by computing its variance, as defined below:

Definition: Variance

Definition: If X is a random variable with mean $\mathbb{E}[X] = \mu$, then the **variance** of X , denoted by $\text{Var}[X] = \sigma^2$, is defined to be

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

Again, this is merely a definition and it does not make sense to prove a definition, but we can certainly take a moment and ask as to why our definition achieves our goal. Recall that our goal is to develop a measure of how much a distribution varies and we are claiming that

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

does the trick. How so?

Well, in order to measure the spread of a distribution, we first must fix a reference point. For us, we are fixing our reference point to be $\mu = \mathbb{E}[X]$ which is the center of our distribution. Once we have μ , we then study $X - \mu$ which represents the distance of each value of X from its center. Since distances cannot be negative, we square the lengths which gives us $(X - \mu)^2$. We then average all the square lengths from each value of X to its center, $\mathbb{E}[(X - \mu)^2]$.

In summary, the expression $\mathbb{E}[(X - \mu)^2]$ yields the average square distance between each value of X and its center.

Some students may argue that if we wish to compute the spread of the distribution, then it makes sense to define the variance as $\mathbb{E}[|X - \mu|]$ since this tells us, on average, how far each value of X is away from its center. Although this is a good idea, in practice, the computations can get messy. This will be further discussed when we study continuous random variables.

✓ Example 9.1.1

Suppose X has the following probability mass function: $f(x) = P(X = x) = \begin{cases} 0.5, & \text{if } x = -1, 1 \\ 0, & \text{otherwise} \end{cases}$

Find $\text{Var}[X]$.

Answer

Recall that for this random variable, $\mu = \mathbb{E}[X] = 0$.

$$\begin{aligned} \text{Var}[X] &= \mathbb{E}[(X - \mu)^2] \\ &= \mathbb{E}[(X - 0)^2] \\ &= \mathbb{E}[X^2] \\ &= \sum_{\text{all } x} x^2 f(x) \\ &= (-1)^2 f(-1) + (1)^2 f(1) \\ &= f(-1) + f(1) \\ &= 0.5 + 0.5 \\ &= 1 \end{aligned}$$

✓ Example 9.1.2

Suppose Y has the following probability mass function: $g(y) = P(Y = y) = \begin{cases} 0.5, & \text{if } y = -100, 100 \\ 0, & \text{otherwise} \end{cases}$

Find $\text{Var}[Y]$.

Answer

Recall that for this random variable, $\mu = \mathbb{E}[Y] = 0$.

$$\begin{aligned} \text{Var}[Y] &= \mathbb{E}[(Y - \mu)^2] \\ &= \mathbb{E}[(Y - 0)^2] \\ &= \mathbb{E}[Y^2] \\ &= \sum_{\text{all } y} y^2 f(y) \\ &= (-10)^2 f(-10) + (10)^2 f(10) \\ &= 100f(-1) + 100f(1) \\ &= 100(0.5) + 100(0.5) \\ &= 100 \end{aligned}$$

As expected, $\text{Var}[Y] > \text{Var}[X]$.

✓ Example 9.1.3

Suppose the probability mass function of X is given by $f(x) = P(X = x) = \begin{cases} .855, & \text{if } x = 0 \\ .140, & \text{if } x = 100000 \\ .005, & \text{if } x = 200000 \\ 0, & \text{otherwise} \end{cases}$

Find $\text{Var}[X]$.

Answer

Recall that for this random variable, $\mu = \mathbb{E}[X] = 15,000$.

$$\begin{aligned}\mathbb{V}ar[X] &= \mathbb{E}\left[(X - \mu)^2\right] \\ &= \mathbb{E}\left[(X - 15,000)^2\right] \\ &= \sum_{\text{all } x} (x - 15000)^2 f(x) \\ &= (0 - 15000)^2 f(0) + (100000 - 15000)^2 f(100000) + (200000 - 15000)^2 f(200000) \\ &= 1,375,000,000\end{aligned}$$

Looking back at the answers to the above three questions, we perhaps may feel uneasy. For instance, in Example 1, the variance is 1 and this makes sense because from the sketch of the pmf,



it makes sense that on average, the square distance of each value of X to its center is 1. For Example 2, on average, the square distance of each value of Y to its center is 100. However, some may argue that the 100 does not make sense since the values of Y are -10 and 10 and so intuitively the most amount of variation we have is 20.

How can we reconcile this?

Simply put, allow us to think about the units of $\mathbb{V}ar[X]$. Notice that in $\mathbb{V}ar[X] = \mathbb{E}\left[(X - \mu)^2\right]$, the $X - \mu$ term is being squared and so our final answer is in terms of $(\text{units of } X)^2$.

For instance, in Example 3, the variance of X is 1,375,000,000 square dollars. However, it is rather difficult in the dimension of dollars squared as opposed to just dollars.

In order to make the measure of spread more intuitive, we will take the square root of the variance. Doing so gives us the standard deviation.

Definition: Standard Deviation

Definition: Given a random variable, X , the positive square root of the variance is called the **standard deviation** of X , denoted by $SD[X]$ or σ .

✓ Example 9.1.4

Suppose X has the following probability mass function: $f(x) = P(X = x) = \begin{cases} 0.5, & \text{if } x = -1, 1 \\ 0, & \text{otherwise} \end{cases}$

Find $SD[X]$.

Answer

Recall that for this random variable, $\mathbb{V}ar[X] = 1$. Hence Find $SD[X] = \sqrt{1} = 1$.

✓ Example 9.1.5

Suppose Y has the following probability mass function: $g(y) = P(Y = y) = \begin{cases} 0.5, & \text{if } y = -100, 100 \\ 0, & \text{otherwise} \end{cases}$

Find $SD[Y]$.

Answer

Recall that for this random variable, $\text{Var}[Y] = 100$. Hence Find $SD[Y] = \sqrt{100} = 10$. Intutively, this means the distribution of Y is ten times more spread out than the distribution of X where X is defined as above in Example 4.

As expected, $\text{Var}[Y] > \text{Var}[X]$.

✓ Example 9.1.6

Suppose the probability mass function of X is given by $f(x) = P(X = x) = \begin{cases} .855, & \text{if } x = 0 \\ .140, & \text{if } x = 100000 \\ .005, & \text{if } x = 200000 \\ 0, & \text{otherwise} \end{cases}$

Find $SD[X]$.

Answer

Recall that for this random variable, $\text{Var}[X] = 1375000000$ Hence Find $SD[X] = \sqrt{1375000000} \approx \$37,081$.

We conclude this section with two theorems concerning the variance. Our first theorem gives us a more convenient way to calculate the variance.

🔗 Theorem 9.1.7

Theorem: If X is a random variable, then

$$\text{Var}[X] = \mathbb{E}[X^2] - \left(\mathbb{E}[X]\right)^2 \quad (9.1.1)$$

That is, the variance is equal to the second moment of X minus the square of the first moment of X . The proof is left as a homework exercise.

🔗 Theorem 9.1.8

Theorem: Let X denote a random variable. For any constants, a and b , $\text{Var}[aX + b] = a^2 \text{Var}[X]$.

9.1: Variances is shared under a [CC BY](#) license and was authored, remixed, and/or curated by LibreTexts.

9.2: The Uniform Random Variable

For the past few weeks, we have been discussing random variables. In this section, we will begin to discuss important random variables and their distributions. Before we do so, allow us to ask a question: *what is the point of a random variable? Or rather, why should we study random variables?*

The answer is the point of a random variable is to model some sort of real life process or phenomenon and this act of modeling a quantity is often the first step towards predicting or forecasting the quantity of interest.

There are certain processes which occurs more frequently than others and so we will attach "brand" names to the random variables which models these processes. Specifically, we will study the Uniform Random Variable, The Bernoulli and Binomial Random Variable, The Geometric and Negative Binomial Random Variable, and the Poisson Random Variable.

We should note that there are infinitely many random variables out there in the universe and so it would be impossible to discuss every such random variable. For that reason, it is nice to have a framework to follow whenever you are introduced to a random variable in your future courses. Here is that framework which we will apply to the Uniform Random Variable, The Bernoulli and Binomial Random Variable, The Geometric and Negative Binomial Random Variable, and the Poisson Random Variable.

Framework for Studying Discrete Random Variables

- 1) We should first ask how is the probability mass function or cumulative distribution function defined.
- 2) We should then realize that the point of a random variable is to model some sort of quantity. Hence, we then ask ourselves what does this random variable models and under what assumptions are required? Essentially, we should try to interpret the random variable.
- 3) When feasible, we should show that the assumptions in (2) yield the probability mass function or cumulative distribution function outlined in (1).

With that said, allow us to apply the framework to our first brand name random variable, The Uniform Random Variable. We first seek to understand how The Uniform Random Variable is defined.

Definition: The (Discrete) Uniform Random Variable

Definition: Let $a \leq b$ be integers. A random variable X whose probability mass function is given by

$$f(x) = P(X = x) = \begin{cases} \frac{1}{b-a+1}, & \text{if } x = a, a+1, a+2, \dots, b \\ 0, & \text{otherwise} \end{cases} \quad (9.2.1)$$

is said to have the Uniform Distribution on the integers $a, a+1, a+2, \dots, b$ and we write $X \sim \mathcal{U}\{a, b\}$ which is read as " X is distributed as a Uniform Random Variable on the integers $a, a+1, a+2, \dots, b$ ".

Again, the point of a random variable is to model some sort of quantity under certain assumptions and so we can ask, "what does the above random variable model and why?" We first answer what does the discrete Uniform Random Variable models.

Interpretation of the (Discrete) Uniform Random Variable

The (discrete) Uniform Random Variable models the outcome of an experiment where the experiment is choosing an integer from $a, \dots, b$ at random.

We now show why the experiment outlined above yields the discrete Uniform Random Variable.

Theorem 9.2.1

Theorem: If X models the outcome of an experiment where the experiment is choosing an integer from $a, \dots, b$ at random, then $X \sim \mathcal{U}\{a, b\}$.

Proof:

To prove this, we simply derive the probability mass function as discussed in [Section 8.1](#). We first ask ourselves what are the possible values of X ? We see that the integer we can select is either $a, a+1, a+2, \dots$ or b and so $X : a, a+1, a+2, \dots, b$. We now must find the probability that X takes each value.

$$\begin{aligned}
 P(X=a) &= P(\underbrace{\{\text{we selected the number } a\}}_{SSS}) = \frac{|\{\text{we selected } a\}|}{|S|} = \frac{1}{b-a+1} \\
 P(X=a+1) &= P(\underbrace{\{\text{we selected the number } a+1\}}_{SSS}) = \frac{|\{\text{we selected } a+1\}|}{|S|} = \frac{1}{b-a+1} \\
 &\dots \\
 &\dots \\
 &\dots \\
 P(X=b) &= P(\underbrace{\{\text{we selected the number } b\}}_{SSS}) = \frac{|\{\text{we selected } b\}|}{|S|} = \frac{1}{b-a+1}
 \end{aligned}$$

Hence, $f(x) = \begin{cases} \frac{1}{b-a+1}, & \text{if } x = a, a+1, a+2, \dots, b \\ 0, & \text{otherwise} \end{cases}$ so $X \sim \mathcal{U}\{a, b\}$.

We end this section with a theorem concerning some of the properties of a discrete Uniform Random Variable.

Theorem 9.2.1

Theorem: If $X \sim \mathcal{U}\{a, b\}$ then:

$$\begin{aligned}
 \mathbb{E}[X] &= \frac{a+b}{2} \\
 \mathbb{V}ar[X] &= \frac{(b-a+1)^2 - 1}{12} \\
 F(x) &= \begin{cases} 0, & \text{if } x < a \\ \frac{|x|-a+1}{b-a+1}, & \text{if } a \leq x < b \\ 1, & \text{if } x \geq b \end{cases}
 \end{aligned}$$

9.2: The Uniform Random Variable is shared under a [CC BY](#) license and was authored, remixed, and/or curated by LibreTexts.

9.3: The Bernoulli and Binomial Random Variables

In this section, we will discuss two other important random variables - the Bernoulli random variable and the Binomial random variable. As usual, we will use our framework for introducing random variables by first defining the pmf or cdf, then by understanding what the random variable models, and then proving our interpretation is correct.

Definition: The Bernoulli Random Variable

Definition: A random variable X whose probability mass function is given by

$$f(x) = P(X = x) = \begin{cases} 1-p & \text{if } x = 0, \\ p & \text{if } x = 1, \\ 0 & \text{otherwise} \end{cases} = \begin{cases} p^x(1-p)^{1-x} & \text{if } x = 0, 1 \\ 0 & \text{otherwise} \end{cases} \quad (9.3.1)$$

is called a Bernoulli Random Variable with parameter p and we write $X \sim \text{Bern}(p)$.

Interpretation of a Bernoulli Random Variable

Suppose an experiment is to be performed with the following characteristics:

1. The experiment can be thought of as consisting of one and only one trial.
2. The trial may result in one of two outcomes where we think of one outcome as being a "success" while the other outcome is labeled as a "failure".
3. The probability of success is p (and hence the probability of failure is $1 - p$).
4. X models the number of successes.

Then $X \sim \text{Bern}(p)$.

To summarize, the Bernoulli random variable models the number of successes when we perform one and only trial.

We now argue that the conditions outlined above yields the Bernoulli random variable.

Theorem: The Bernoulli Random Variable 9.3.1

Theorem: Suppose an experiment is to be performed with the following characteristics:

1. The experiment consists of one and only one trial.
2. The trial may result in either a success or a failure.
3. The probability of success is p and the probability of failure is $1 - p$.
4. X models the number of successes.

Then $X \sim \text{Bern}(p)$.

Proof:

We now derive the probability mass function of X . First consider the possible values of X . Suppose we are conducting one and only one trial. By assumption, this trial may result either in a success or a failure. If we have a success, then the number of successes is equal to 1 and so $X = 1$. On the other hand, if the trial results in a failure, then the number of successes we have is 0 and so $X = 0$. Thus, $X : 0, 1$. Allow us to now find the probability associated to each value.

$P(X = 0) = P(\text{failure}) = 1 - p$ and similarly, $P(X = 1) = P(\text{success}) = p$. Hence,

$$f(x) = P(X = x) = \begin{cases} 1-p & \text{if } x = 0, \\ p & \text{if } x = 1, \\ 0 & \text{otherwise} \end{cases} = \begin{cases} p^x(1-p)^{1-x} & \text{if } x = 0, 1 \\ 0 & \text{otherwise} \end{cases}$$

which completes the derivation.

Example 9.3.2

Suppose that the probability that a single seed will germinate is 0.8. If X models the number of successes, then X is a Bernoulli Random Variable with parameter 0.8 because

1. There is one and only trial (where for us, a trial is planting a seed).
2. That trial can either result in one of two possible outcomes. We can think of the seed germinating as a success and the seed not germinating as a failure.
3. The probability of success is 0.8 and the probability of failure is 0.2.
4. X models the number of successes.

Hence, we write $X \sim \text{Bern}(0.8)$. That is,

$$f(x) = P(X = x) = \begin{cases} 0.2 & \text{if } x = 0, \\ 0.8 & \text{if } x = 1, \\ 0 & \text{otherwise} \end{cases}$$

✓ Example 9.3.3

Suppose an experiment involves drawing a single card from a deck of 52. If X models the number of aces drawn then $X \sim \text{Bern}(4/52)$. That is,

$$f(x) = P(X = x) = \begin{cases} 48/52 & \text{if } x = 0, \\ 4/52 & \text{if } x = 1, \\ 0 & \text{otherwise} \end{cases}$$

✓ Example 9.3.4

Suppose a company produces light bulbs and there is a 1% chance any light bulb will be defective. If we buy a single light bulb, and X models the number of defective light bulbs, then $X \sim \text{Bern}(0.01)$. That is,

$$f(x) = P(X = x) = \begin{cases} 0.99 & \text{if } x = 0, \\ 0.01 & \text{if } x = 1, \\ 0 & \text{otherwise} \end{cases}$$

✓ Example 9.3.5

Suppose an experiment involves flipping a fair coin once. If X models the number of heads obtained, then find the distribution of X .

Answer

$$X \sim \text{Bern}(0.5).$$

✓ Example 9.3.6

Suppose a medical trial consists of administering a drug to a single patient and then seeing if the drug is a success (meaning beneficial) for that patient or if the drug is a failure (meaning non-beneficial) for that patient. The probability the drug is successful for any patient is 0.75. Let X model the number of patients for which the drug is successful. Find the distribution of X .

Answer

$$X \sim \text{Bern}(0.75).$$

We now present the expected value, variance, and cumulative distribution function for the Bernoulli random variable.

Theorem 9.3.7

Theorem: If $X \sim \text{Bern}(p)$ then

$$\begin{aligned} \mathbb{E}[X] &= p \\ \text{Var}[X] &= p(1 - p) \end{aligned}$$

Additionally, the cumulative distribution function is given by

$$F(x) = P(X \leq x) = \begin{cases} 0 & \text{if } x < 0, \\ 1 - p & \text{if } 0 \leq x < 1, \\ 1 & \text{if } x \geq 1 \end{cases} \quad (9.3.2)$$

Allow us to modify Example 9.3.5. What if instead of administering the drug to one and only patient and modeling the number of resulting successes, we instead administered the drug to five patients and wished to model the number of successes. What would this look like? More generally, we are asking if instead of having one trial, we performed many independent trials, say n of them and wished to model the number of successes by a random variable X . In such a case, what would the probability mass function of X look like? Clearly X is no longer Bernoulli since it is possible for the number of successes to be 2 or 3 or $\dots n$.

It would be nice to have a random variable which models the number of successes among n independent trials. The above scenario calls on us to use another brand name random variable which an extension of the Bernoulli. This random variable is the Binomial random variable.

Definition: The Binomial Random Variable

Definition: A random variable X whose probability mass function is given by

$$f(x) = P(X = x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & \text{if } x = 0, 1, 2, \dots, n \\ 0 & \text{otherwise} \end{cases} \quad (9.3.3)$$

is called a Binomial random variable with parameters n and p and we write $X \sim \text{Bin}(n, p)$.

Interpretation of a Bernoulli Random Variable

Suppose an experiment is to be performed with the following characteristics:

1. The experiment can be thought of as consisting of n independent trials.
2. The trial may result in one of two outcomes where we think of one outcome as being a "success" while the other outcome is labeled as a "failure".
3. The probability of success is p (and hence the probability of failure is $1 - p$).
4. X models the total number of successes.

Then $X \sim \text{Bin}(n, p)$.

To summarize, the Binomial random variable models the total number of successes among n independent trials.

We now argue that the conditions outlined above yields the Bernoulli random variable.

Theorem: The Bernoulli Random Variable 9.3.1

Theorem: Suppose an experiment is to be performed with the following characteristics:

1. The experiment can be thought of as consisting of n independent trials.
2. The trial may result in one of two outcomes where we think of one outcome as being a "success" while the other outcome is labeled as a "failure".
3. The probability of success is p (and hence the probability of failure is $1 - p$).
4. X models the total number of successes.

Then $X \sim \text{Bin}(n, p)$.

Proof:

We already did this! This was answered as a [homework exercise](#).

Example 9.3.8

Allow us to reconsider the following [example](#) from a previous section: a sequence of five independent trials are to be performed. Each trial consists of administering a drug to a patient and then seeing if the drug is a success (meaning beneficial) for that patient or if the drug is a failure (meaning non-beneficial) for that patient. The probability the drug is successful for any patient is 0.75. Find the probability that the drug is successful for at least one patient.

Answer

We will answer this question from the point of view of random variables. In this question, we are interested in modeling the number of successes that occurs and so let X model the number of successes among the five independent trials. Notice that

1. The experiment can be thought of as consisting of 5 independent trials.
2. Each trial either results in a success or failure.
3. The probability of success on any trial is 0.75 (and hence the probability of failure of 0.25).
4. X models the total number of successes.

Then $X \sim \text{Bin}(5, 0.75)$. That is,

$$f(x) = P(X = x) = \begin{cases} \binom{5}{x} 0.75^x (0.25)^{5-x} & \text{if } x = 0, 1, 2, 3, 4, 5 \\ 0 & \text{otherwise} \end{cases} \quad (9.3.4)$$

Then

$$P(\text{drug is successful for at least one patient}) = P(X \geq 1) = 1 - P(X < 1) = 1 - P(X = 0) = 1 - f(0) = 1 - \binom{5}{0} 0.75^0 0.25^5 = 1 - 0.25^5 \approx 0.9990$$

Note

Knowing the appropriate calculator commands will be useful for future problems. Suppose $X \sim \text{Bin}(n, p)$ and let k be in the support of X . Then

1. $f(k) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} = \text{binompdf}(n, p, k)$.
2. $F(k) = P(X \leq k) = \binom{n}{0} p^0 (1-p)^n + \binom{n}{1} p^1 (1-p)^{n-1} + \binom{n}{2} p^2 (1-p)^{n-2} + \dots + \binom{n}{k} p^k (1-p)^{n-k} = \text{binomcdf}(n, p, k)$

Note that we click 2^{nd} followed by Vars/Distribution on your calculator and scroll down to the appropriate command.

Hence, $P(\text{drug is successful for at least one patient}) = P(X \geq 1) = 1 - P(X < 1) = 1 - P(X = 0) = 1 - \text{binompdf}(5, 0.75, 0) \approx 0.9990$.

✓ Example 9.3.9

Allow us to reconsider the following [example](#) from a previous section: a sequence of five independent trials are to be performed. Each trial consists of administering a drug to a patient and then seeing if the drug is a success (meaning beneficial) for that patient or if the drug is a failure (meaning non-beneficial) for that patient. The probability the drug is successful for any patient is 0.75. Find the probability that exactly three successes occur.

Answer

We will answer this question from the point of view of random variables. In this question, we are interested in modeling the number of successes that occurs and so let X model the number of successes among the five independent trials. Then $X \sim \text{Bin}(5, 0.75)$. That is,

$$f(x) = P(X = x) = \begin{cases} \binom{5}{x} 0.75^x (0.25)^{5-x} & \text{if } x = 0, 1, 2, 3, 4, 5 \\ 0 & \text{otherwise} \end{cases} \quad (9.3.5)$$

Then $P(\text{drug is successful for three patients}) = P(X = 3) = f(3) = \binom{5}{3} 0.75^3 0.25^2 \approx 0.2637$ or alternatively, $P(\text{drug is successful for three patients}) = P(X = 3) = \text{binompdf}(5, 0.75, 3) \approx 0.2637$.

✓ Example 9.3.10

Suppose an experiment involves planting ten seeds and checking to see how many seeds germinate. According to the manufacturer, there is a 80% chance a seed will germinate. If the growth of the seeds are independent of each other, find the probability that exactly 7 seeds will germinate.

Answer

Let X model the number of seeds that germinate. Notice that,

1. The experiment can be thought of as consisting of 10 independent trials where a trial is planting a seed.
2. That trial can either result in one of two possible outcomes. We can think of the seed germinating as a success and the seed not germinating as a failure.
3. The probability of success is 0.8 and the probability of failure is 0.2.
4. X models the number of successes.

Hence, $X \sim \text{Bin}(10, 0.8)$. That is,

$$f(x) = P(X = x) = \begin{cases} \binom{10}{x} 0.8^x (0.2)^{10-x} & \text{if } x = 0, 1, 2, 3, \dots, 10 \\ 0 & \text{otherwise} \end{cases} \quad (9.3.6)$$

Hence $P(\text{exactly 7 seeds will germinate}) = P(X = 7) = f(7) = \binom{10}{7} 0.8^7 (0.2)^3 \approx 0.2013 = 20.13\%$ or alternatively $P(\text{exactly 7 seeds will germinate}) = P(X = 7) = \text{binompdf}(10, 0.8, 7) \approx 0.2013 = 20.13\%$

✓ Example 9.3.11

It is known that light bulbs produced by a certain company will be defective with probability 0.01 independently of one another. The company sells the light bulbs in packs of 15 and states at most, one of the fifteen bulbs will be defective. If more than 1 bulb is defective, then the company offers a money back guarantee. Suppose you buy a pack of light bulbs. What is the probability that your pack will be eligible for the money back guarantee?

Answer

Let X model the number of defective light bulbs. In order for the pack to be eligible for the money back guarantee, it means that we have more than 1 defective light bulb. So we must find $P(X > 1)$.

- 1) We have a total of 15 independent trials where we can think of each trial as examining each light bulb and checking to see if it is defective or working.

- 2) Each trial either results in a success (defective) or failure (working).
- 3) The probability of success on each trial is 0.01 and the probability of failure on each trial is 0.99.
- 4) X models the total number of successes.

Hence $X \sim \text{Bin}(15, 0.01)$ That is,

$$f(x) = P(X = x) = \begin{cases} \binom{15}{x} 0.01^x (0.99)^{15-x} & \text{if } x = 0, 1, 2, 3, \dots, 15 \\ 0 & \text{otherwise} \end{cases} \quad (9.3.7)$$

So

$$\begin{aligned} P(X > 1) &= P(X = 2) + P(X = 3) + P(X = 4) + \dots + P(X = 15) \\ &= 1 - P(X \leq 1) \\ &= 1 - [P(\{X = 0\} \cup \{X = 1\})] \\ &= 1 - [P(X = 0) + P(X = 1)] \\ &= 1 - \left[\binom{15}{0} 0.01^0 (0.99)^{15} + \binom{15}{1} 0.01^1 (0.99)^{14} \right] \\ &= 0.0096297734 \end{aligned}$$

Alternatively, we can say that

$$\begin{aligned} P(X > 1) &= P(X = 2) + P(X = 3) + P(X = 4) + \dots + P(X = 15) \\ &= 1 - P(X \leq 1) \\ &= 1 - [P(\{X = 0\} \cup \{X = 1\})] \\ &= 1 - [\text{binompdf}(15, 0.01, 0) + \text{binompdf}(15, 0.01, 1)] \\ &= 0.0096297734 \end{aligned}$$

or even more succinctly,

```
begin{align*}
P(X > 1) &= P(X=2) + P(X=3) + P(X=4) + \ldots + P(X=15) \\
&= 1 - P(X \leq 1) \\
&= 1 - [P(\text{binomcdf}(15, 0.01, 1))] \\
&= 0.0096297734
\end{align*}
```

We now present the expected value and variance for the Bernoulli random variable.

Theorem 9.3.12

Theorem: If $X \sim \text{Bin}(n, p)$ then

$$\begin{aligned} \mathbb{E}[X] &= np \\ \text{Var}[X] &= np(1 - p) \end{aligned}$$

9.3: The Bernoulli and Binomial Random Variables is shared under a CC BY license and was authored, remixed, and/or curated by LibreTexts.

9.4: The Geometric and Negative Binomial Random Variables

In this section, we will discuss another two important random variables - the Geometric random variable and the Negative Binomial random variable. As usual, we will use our framework for introducing random variables by first defining the pmf or cdf, then by understanding what the random variable models, and then proving our interpretation is correct.

Definition: The Geometric Random Variable

Definition: A random variable X whose probability mass function is given by

$$f(x) = P(X = x) = \begin{cases} p(1-p)^{x-1} & \text{if } x = 1, 2, 3, \dots \\ 0 & \text{otherwise} \end{cases} \quad (9.4.1)$$

is called a Geometric random variable with parameter p and we write $X \sim \text{Geo}(p)$.

Interpretation of a Geometric Random Variable

Suppose an experiment is to be performed with the following characteristics:

1. The experiment can be thought of as repeating independent trials until we reach our first success at which point the experiment stops.
2. The trial may result in one of two outcomes where we think of one outcome as being a "success" while the other outcome is labeled as a "failure".
3. The probability of success on every trial is p (and hence the probability of failure is $1 - p$).
4. X models the number of trials it takes to obtain our first success.

Then $X \sim \text{Geo}(p)$.

To summarize, the Geometric random variable models the number of trials it takes to obtain our first success.

Note: There are actually two different ways to define the Geometric random variable. Some sources define the Geometric random variable as modeling the number of trials until the first success which is what we have done. Other sources may define the Geometric random variable to be the number of failures until the first success. Thus, if you look at other resources, make sure you know which definition of a Geometric random variable they are using! (As you can see, Wikipedia lists both: https://en.Wikipedia.org/wiki/Geometric_distribution).

We now argue that the conditions outlined above yields the Geometric random variable.

Theorem: The Geometric Random Variable 9.4.1

Theorem: Suppose an experiment is to be performed with the following characteristics:

1. The experiment can be thought of as repeating independent trials until we reach our first success at which point the experiment stops.
2. The trial may result in one of two outcomes where we think of one outcome as being a "success" while the other outcome is labeled as a "failure".
3. The probability of success is p (and hence the probability of failure is $1 - p$).
4. X models the number of trials it takes to obtain our first success.

Then $X \sim \text{Geo}(p)$.

Proof:

We now derive the probability mass function of X . First consider the possible values of X . Remember that X denotes the number of trials until our first success. Perhaps on our first trial we obtained our first success and so $X = 1$. Or perhaps on our second trial we obtained our first success. This means we must have failed first and then succeeded. In this case, $X = 2$. Or perhaps on our third trial we obtained our first success. In terms of the sequence of trials, this means we failed, then failed, then succeeded and so $X = 3$ and so on and so forth. Thus, the possible values of X are $1, 2, 3, \dots$.

Allow us to now find the probability that X takes each of these values. Since there are infinitely many values, we hope to find some sort of general pattern. For convenience, let $\{S_i = \{\text{we have a success on trial } i\}\}$ and let $\{F_i = \{\text{we have a failure on trial } i\}\}$. Then

1. $P(X = 1) = P(S_1) = p$
2. $P(X = 2) = P(F_1 \cap S_2) = P(F_1)P(S_2) = (1-p)p = p(1-p)$
3. $P(X = 3) = P(F_1 \cap F_2 \cap S_3) = P(F_1)P(F_2)P(S_3) = (1-p)(1-p)p = p(1-p)^2$
4. $P(X = 4) = P(F_1 \cap F_2 \cap F_3 \cap S_4) = P(F_1)P(F_2)P(F_3)P(S_4) = (1-p)(1-p)(1-p)p = p(1-p)^3$

And more generally, we see for any x in the support of X , we have the following:

$$\begin{aligned} P(X = x) &= P(\text{the first success occurs on the } x^{\text{th}} \text{ trial}) = P(F_1 \cap F_2 \cap \dots \cap F_{x-1} \cap S_x) = P(F_1)P(F_2) \dots P(F_{x-1})P(S_x) = (1-p)^{x-1}p \\ &= p(1-p)^{x-1} \end{aligned}$$

Hence,

$$f(x) = P(X = x) = \begin{cases} p(1-p)^{x-1} & \text{if } x = 1, 2, 3, \dots \\ 0 & \text{otherwise} \end{cases}$$

which completes the derivation.

Note

For the next example, recall from calculus that

$$a + ar + ar^2 + ar^3 + \dots = \sum_{n=1}^{\infty} ar^{n-1} = \frac{a}{1-r}, \quad |r| < 1$$

✓ Example 9.4.2

Reconsider the following [example](#): Suppose an experiment consists of repeatedly flipping a fair coin until a heads appears for the first time at which point the experiment stops. Assuming that the flips of the coin are independent of each other, find the probability that the experiment will eventually terminate. That is, find the probability that a heads is eventually obtained.

Answer

Let X denote the number of trials it takes to obtain our first heads. Notice that the experiment can be thought of as repeatedly performing independent flips of a fair coin until we obtain a heads, each trial (flip) results either in a heads (success) or failure (tails), the probability of success on each trial is 0.5, and X models the number of trials until our first success. Hence, $X \sim \text{Geo}(0.5)$. That is,

$$f(x) = P(X = x) = \begin{cases} 0.5(1-0.5)^{x-1} & \text{if } x = 1, 2, 3, \dots \\ 0 & \text{otherwise} \end{cases}$$

We can simplify the above to

$$f(x) = P(X = x) = \begin{cases} (0.5)^x & \text{if } x = 1, 2, 3, \dots \\ 0 & \text{otherwise} \end{cases}$$

We wish to find the probability that a heads is eventually obtained. That is, we could have obtained our first heads on trial 1 or trial 2 or trial 3 or so on and so fourth. Hence, we obtain the following:

$$\begin{aligned} P(\text{we eventually obtain a heads}) &= P(\text{our first heads appears on trial 1 or trial 2 or trial 3 or } \dots) \\ &= P(\{X = 1\} \cup \{X = 2\} \cup \{X = 3\} \cup \dots) \\ &= P(X = 1) + P(X = 2) + P(X = 3) + \dots \\ &= 0.5^1 + 0.5^2 + 0.5^3 + \dots \\ &= \frac{0.5}{1-0.5} \\ &= 1 \end{aligned}$$

Of course, this answer agrees with our computation from Section 6.2. Further note that the computation is essentially $\sum_{\text{all } x} f(x)$ and by the properties discussed of the probability mass function, this summation is equal to 1.

Theorem 9.4.3

Theorem: If $X \sim \text{Geo}(p)$, then for any integer k in the support of X , $P(X > k) = (1-p)^k$.

Proof:

$$\begin{aligned} P(X > k) &= P(X = k+1) + P(X = k+2) + P(X = k+3) \dots \\ &= p(1-p)^k + p(1-p)^{k+1} + p(1-p)^{k+2} \dots \\ &= \frac{p(1-p)^k}{1-(1-p)} \\ &= \frac{p(1-p)^k}{p} \\ &= (1-p)^k \end{aligned}$$

Alternatively, notice the interpretation. Since $X > k$, then the number of trials to reach our first success is greater than k . This happens if and only if the first k trials are all failures and so $P(X > k) = P(\text{first } k \text{ trials are all failures}) = P(F_1 \cap F_2 \cap \dots \cap F_k) = P(F_1)P(F_2) \dots P(F_k) = (1-p)^k$.

We now present the expected value, variance, and cumulative distribution function for the Geometric random variable.

Theorem 9.4.4

Theorem: If $X \sim \text{Geo}(p)$ then

$$\mathbb{E}[X] = \frac{1}{p}$$

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

Additionally, the cumulative distribution function is given by

$$F(x) = P(X \leq x) = \begin{cases} 0 & \text{if } x < 1, \\ 1 - (1-p)^{\lfloor x \rfloor} & \text{if } x \geq 1 \end{cases} \quad (9.4.2)$$

✓ Example 9.4.1

An experiment consists of administering a drug to a patient and then seeing if the drug is a success (meaning beneficial) for that patient or if the drug is a failure (meaning non-beneficial) for that patient. Let us suppose that a sequence of independent trials are to be performed until a person is found for which the drug is successful. It is known that the probability the drug is successful for any patient is 0.75. Find the probability that it takes more than five patients to obtain our first success.

Answer

We are interested in modeling the number of patients it takes to obtain our first success and so allow us to define our random variable. Let X denote the number of patients it takes to obtain our first success. Notice that:

- 1) The experiment can be thought of as repeating independent trials until we reach our first success at which point the experiment stops.
- 2) The trial may result in one of two outcomes where we think of one outcome as being a "success" while the other outcome is labeled as a "failure".
- 3) The probability of success on every trial is 0.75 (and hence the probability of failure is 0.25).
- 4) X models the number of trials it takes to obtain our first success.

Hence, $X \sim \text{Geo}(0.75)$. Thus, by Theorem 9.4.3 we have the following:

$$P(\text{it takes more than five patients to obtain our first success}) = P(X > 5) = (1 - 0.75)^5 = 0.25^5$$

Alternatively, by Theorem 9.3.4 we have the following:

$$P(\text{it takes more than five patients to obtain our first success}) = P(X > 5) = 1 - P(X \leq 5) = 1 - [1 - (1 - 0.75)^5] = 0.25^5$$

and finally, we may also write the following:

we have the following:

$$P(\text{it takes more than five patients to obtain our first success}) = P(X > 5) = 1 - P(X \leq 5) = 1 - \text{geometcdf}(0.75, 5)$$

✓ Example 9.4.1

An experiment consists of administering a drug to a patient and then seeing if the drug is a success (meaning beneficial) for that patient or if the drug is a failure (meaning non-beneficial) for that patient. Let us suppose that a sequence of independent trials are to be performed until a person is found for which the drug is successful. It is known that the probability the drug is successful for any patient is 0.75. Find the probability that it takes at most three patients to obtain our first success.

Answer

Similarly, let X denote the number of patients it takes to obtain our first success. As argued above, $(X \sim \text{Geo}(0.75))$. Thus, by Theorem 9.4.4 we have the following:

$$P(\text{it takes at most three patients to obtain our first success}) = P(X \leq 3) = [1 - (1 - 0.75)^3] = 1 - 0.25^3$$

Alternatively, we may use the following calculator command:

$$P(\text{it takes at most three patients to obtain our first success}) = P(X \leq 3) = \text{geometcdf}(0.75, 3)$$

✓ Example 9.4.1

It is known that light bulbs produced by a certain company will be defective with probability 0.01 independently of one another. The company sells the light bulbs in packs of 15 and states at most, one of the fifteen bulbs will be defective. If more than 1 bulb is defective, then the company offers a money back guarantee. Suppose an experiment consists of a customer who continually buys packs until they come across a pack that is eligible for return. What is the probability that the customer will have to buy more than 20 packs?

Answer

Let X model the number of packs the customer buys until they come across a pack that is eligible for return. Then $X \sim Geo(p)$. Can we find p ? That is, can we find what is the probability that the customer will return a pack?

Well the customer returning a pack depends on the number of defective light bulbs. Let Y denote the number of defective light bulbs. Then $Y \sim Bin(15, 0.01)$.

$$p = P(\text{a pack is eligible for return}) = P(Y > 1) = 1 - P(Y \leq 1) = 1 - \text{binomcdf}(15, 0.01, 1) \approx 0.0096297734$$

Hence $X \sim Geo(0.0096297734)$

$$\begin{aligned} P(\text{the customer has to buy more than 20 packs until they find a pack eligible for return}) &= P(X > 20) \\ &= (1 - 0.0096297734)^{20} \\ &= 0.8240461109 \end{aligned}$$

Or alternatively,

$$\begin{aligned} P(\text{the customer has to buy more than 20 packs until they find a pack eligible for return}) &= P(X > 20) \\ &= 1 - P(X \leq 20) \\ &= 1 - \text{geometcdf}(0.0096297734, 20) \\ &= 0.8240461109 \end{aligned}$$

Allow us to modify Example 9.3.5. What if instead of modeling the number of trials until our first success, we wish to model the number of trials until our second success. What would this look like? More generally, we are asking if instead of repeating trials until our first success, we repeated trials until we obtained a total of r successes and wished to model the number of trials until our r^{th} success by a random variable X . In such a case, what would the probability mass function of X look like? Clearly X is no longer Geometric (provided that $r > 1$).

It would be nice to have a random variable which models the number of trials it takes to obtain our r^{th} success. The above scenario calls on us to use another brand name random variable which is an extension of the Geometric. This random variable is the Negative Binomial random variable.

Definition: The Negative Binomial Random Variable

Definition: A random variable X whose probability mass function is given by

$$f(x) = P(X = x) = \begin{cases} \binom{x-1}{r-1} p^r (1-p)^{x-r} & \text{if } x = r, r+1, r+2, \dots \\ 0 & \text{otherwise} \end{cases} \quad (9.4.3)$$

is called a Negative Binomial random variable with parameters r and p and we write $X \sim NegBin(r, p)$.

Notice that $NegBin(1, p) = Geo(p)$.

Interpretation of a Negative Binomial Random Variable

Suppose an experiment is to be performed with the following characteristics:

1. The experiment can be thought of as repeating independent trials until we reach our r^{th} success at which point the experiment stops.
2. The trial may result in one of two outcomes where we think of one outcome as being a "success" while the other outcome is labeled as a "failure".
3. The probability of success on every trial is p (and hence the probability of failure is $1 - p$).
4. X models the number of trials it takes to obtain our r^{th} success.

Then $X \sim NegBin(r, p)$.

To summarize, the Negative Binomial random variable models the total number of independent trials it takes to obtain r successes.

Theorem: The Negative Binomial Random Variable 9.4.1

Theorem: Suppose an experiment is to be performed with the following characteristics:

1. The experiment can be thought of as repeating independent trials until we reach our r^{th} success at which point the experiment stops.
2. The trial may result in one of two outcomes where we think of one outcome as being a "success" while the other outcome is labeled as a "failure".
3. The probability of success on every trial is p (and hence the probability of failure is $1 - p$).

4. X models the number of trials it takes to obtain our r^{th} success.

Then $X \sim \text{NegBin}(r, p)$.

Proof:

Left as a homework exercise. Hint: If the r^{th} success occurs on the x^{th} trial, then what can we say occurred on the previous $x - 1$ trials?

✓ Example 9.4.10

Suppose an experiment consists of administering a drug to a patient and then seeing if the drug is a success (meaning beneficial) for that patient or if the drug is a failure (meaning non-beneficial) for that patient. Let us suppose that a sequence of independent trials are to be performed until ten people are found for which the drug is successful. It is known that the probability the drug is successful for any patient is 0.75. Find the probability that the tenth success occurs on the thirtieth patient.

Answer

Let X model the number of patients it takes to obtain our tenth success. Notice that

- 1) The experiment can be thought of as repeating independent trials until we reach our tenth success at which point the experiment stops.
- 2) The trial may result in one of two outcomes where we think of one outcome as being a "success" while the other outcome is labeled as a "failure".
- 3) The probability of success on every trial is 0.75 (and hence the probability of failure is 0.25).
- 4) X models the number of trials it takes to obtain our tenth success.

Hence, $X \sim \text{NegBin}(10, 0.75)$ That is,

$$f(x) = P(X = x) = \begin{cases} \binom{x-1}{9} 0.75^{10} (0.25)^{x-10} & \text{if } x = 10, 11, 12, \dots \\ 0 & \text{otherwise} \end{cases}$$

Hence $P(\text{the tenth success occurs on the thirtieth patient}) = P(X = 30) = f(30) = \binom{29}{9} 0.75^{10} (0.25)^{20}$ or alternatively, we may write $P(\text{the tenth success occurs on the thirtieth patient}) = P(X = 30) = f(30) = \text{NegBinpdf}(10, 0.75, 30)$

✓ Example 9.4.10

Suppose an experiment consists of administering a drug to a patient and then seeing if the drug is a success (meaning beneficial) for that patient or if the drug is a failure (meaning non-beneficial) for that patient. Let us suppose that a sequence of independent trials are to be performed until ten people are found for which the drug is successful. It is known that the probability the drug is successful for any patient is 0.75. Find the probability that it takes more than thirty patients to obtain our tenth success.

Answer

Let X model the number of patients it takes to obtain our tenth success. As argued above, $(X \sim \text{NegBin}(10, 0.75))$. That is,

$$f(x) = P(X = x) = \begin{cases} \binom{x-1}{9} 0.75^{10} (0.25)^{x-10} & \text{if } x = 10, 11, 12, \dots \\ 0 & \text{otherwise} \end{cases}$$

Hence

$P(\text{it takes more than thirty patients to obtain our tenth success}) = P(X > 30) = 1 - P(X \leq 30) = 1 - \text{NegBincdf}(10, 0.75, 30)$.

We now present the expected value and variance for the Negative Binomial random variable which concludes this section.

Theorem 9.4.12

Theorem: If $X \sim \text{NegBin}(r, p)$ then

$$\begin{aligned} \mathbb{E}[X] &= \frac{r}{p} \\ \text{Var}[X] &= \frac{r(1-p)}{p^2} \end{aligned}$$

9.4: The Geometric and Negative Binomial Random Variables is shared under a CC BY license and was authored, remixed, and/or curated by LibreTexts.

9.5: The Poisson Random Variable

In this section, we will discuss our last important discrete random variable - the Poisson random variable. As usual, we will use our framework for introducing random variables by first defining the pmf or cdf, then by understanding what the random variable models, and then proving our interpretation is correct (when feasible).

Definition: The Geometric Random Variable

Definition: Let λ denote a positive real number. A random variable X whose probability mass function is given by

$$f(x) = P(X = x) = \begin{cases} e^{-\lambda} \times \frac{\lambda^x}{x!} & \text{if } x = 0, 1, 2, 3, \dots \\ 0 & \text{otherwise} \end{cases} \quad (9.5.1)$$

is called a Poisson random variable with parameter λ and we write $X \sim \text{Poisson}(\lambda)$.

We offer two interpretations of a Poisson random variable.

Interpretation 1 of a Poisson Random Variable

Suppose an experiment is to be performed with the following characteristics:

1. The random variable X can be thought of as a Binomial random variable with parameters n and p .
2. n is sufficiently large.
3. p is sufficiently small.

Then $X \sim \text{Poisson}(\lambda)$ where $\lambda = np$.

Interpretation 2 of a Poisson Random Variable

Suppose an experiment is to be performed with the following characteristics:

1. $X(t)$ counts or models the number of events which occurs by time t .
2. $X(0) = 0$.
3. $\{X(t), t \geq 0\}$ has independent increments meaning the number of events that occurs in disjoint (nonoverlapping) time intervals are independent.
4. $P(X(t+h) - X(t) = 1) = \lambda h + o(h)$ where the function f is said to be $o(h)$ if $\lim_{h \rightarrow 0} \frac{f(h)}{h} = 0$.
5. $P(X(t+h) - X(t) \geq 2) = o(h)$.

Then $X \sim \text{Poisson}(\lambda)$.

Essentially, Interpretation 1 is positing that the Poisson random variable can be used as an approximation for a Binomial random variable when n is large and p is small enough. Meanwhile, Interpretation 2 is stating that the Poisson random variable models the number of times a particular event occurs under certain conditions.

Our focus will be on Interpretation 1 and so allow us to give an example of when we have a Poisson distribution. For instance, if X models the number of customers arriving at a store, then we can argue that X is a Poisson random variable. To argue this, suppose we know that customers arrive to a store, on average, 10 customers per hour. What we can do is divide the hour into minutes or seconds or milliseconds and consider each time unit as a trial. If we were to divide the hour into seconds, then we have 3,600 seconds and hence 3,600 trials. On any given second/trial, the probability that a customer will arrive to the store is $\frac{10}{3600}$ and hence we have a Binomial process with a large value of $n = 3600$ and a small value of $p = \frac{10}{3600}$. We will now prove the validity of Interpretation 1.

Theorem: The Poisson Random Variable 9.5.1

Theorem: Suppose an experiment is to be performed with the following characteristics:

1. The random variable X can be thought of as a Binomial random variable with parameters n and p .
2. n is sufficiently large.
3. p is sufficiently small.

Then $X \sim \text{Poisson}(\lambda)$ where $\lambda = np$.

Proof:

We now derive the probability mass function of X . First consider the possible values of X . Remember that X denotes the number of occurrences of an event and so $x = 0, 1, 2, 3, \dots$. Thus, the possible values of X are $0, 1, 2, 3, \dots$.

We will now find the probability that X takes each of these values. Allow us to argue this generally by finding an approximation for $P(X = x)$

where x is some value in the support of X . Since X is a Binomial random variable, $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$. Define $\lambda = np$

and so $p = \frac{\lambda}{n}$. Doing so, we obtain the following:

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

Substituting in $p = \frac{\lambda}{n}$

$$= \binom{n}{x} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x}$$

Expanding the binomial

$$= \frac{n!}{x!(n-x)!} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x}$$

Simplifying the $\frac{n!}{(n-x)!}$ yields

$$= \frac{n(n-1)(n-2)\dots(n-x+1)}{x!} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x}$$

Rewriting the last two terms yields

$$= \frac{n(n-1)(n-2)\dots(n-x+1)}{x!} \frac{\lambda^x (1 - \frac{\lambda}{n})^n}{n^x (1 - \frac{\lambda}{n})^x}$$

Rearranging some of the terms yields

$$= \frac{n(n-1)(n-2)\dots(n-x+1)}{n^x} \frac{\lambda^x (1 - \frac{\lambda}{n})^n}{x! (1 - \frac{\lambda}{n})^x}$$

For a sufficiently large n , notice that

$$\frac{n(n-1)(n-2)\dots(n-x+1)}{n^x} = \frac{n(n-1)(n-2)\dots(n-x+1)}{n \times n \times \dots \times n}$$

$$= \frac{n}{n} \times \frac{n-1}{n} \times \frac{n-2}{n} \times \dots \times \frac{n-x+1}{n}$$

$$= 1 \times \left(1 - \frac{1}{n}\right) \times \left(1 - \frac{2}{n}\right) \times \dots \times \left(1 - \frac{x-1}{n}\right) \rightarrow 1 \times (1-0) \times (1-0) \times \dots \times (1-0) = 1$$

Also recall from calculus that for a sufficiently large n , we have $\left(1 - \frac{\lambda}{n}\right)^n \rightarrow e^{-\lambda}$ and $\left(1 - \frac{\lambda}{n}\right)^x \rightarrow 1$.

Putting everything together, we see that

$$\begin{aligned} P(X = x) &= \frac{n(n-1)(n-2)\dots(n-x+1)}{\underbrace{n^x}_1} \frac{\lambda^x (1 - \frac{\lambda}{n})^n}{x! \underbrace{(1 - \frac{\lambda}{n})^x}_{e^{-\lambda}}} \\ &\rightarrow \frac{\lambda^x}{x!} e^{-\lambda} \\ &= e^{-\lambda} \frac{\lambda^x}{x!} \end{aligned}$$

Allow us to consider an example which shows the relationship between the Binomial and the Poisson.

✓ Example 9.5.2

A manufacturer of light bulbs knows that 2% of it's bulbs are defective. The manufacturer sells the bulbs in packs of 100. Assuming that the bulbs are independent of each other, find the probability that the pack contains at most three defective bulbs.

Answer

Let X denote the random variable which models the number of defective bulbs in a pack. Then $X \sim \text{Bin}(100, 0.02)$ Hence,

$$P(\text{we obtain at most three defective bulbs}) = P(X \leq 3) = \text{binomcdf}(100, 0.02, 3) \approx 0.85896$$

✓ Example 9.5.3

A manufacturer of light bulbs knows that 2% of it's bulbs are defective. The manufacturer sells the bulbs in packs of 100. Assuming that the bulbs are independent of each other, find the probability that the pack contains at most three defective bulbs.

Answer

Let X denote the random variable which models the number of defective bulbs in a pack. Then $X \sim \text{Bin}(100, 0.02)$ However, we may argue that $n = 100$ is sufficiently large and that $p = 0.02$ is sufficiently small. In doing so, we are saying that $X \sim \text{Poisson}(\lambda)$ where $\lambda = np = 100(0.02) = 2$ and so X is approximately distributed as a Poisson(2) random variable. Hence,

$$P(\text{we obtain at most three defective bulbs}) = P(X \leq 3) \approx \text{poissoncdf}(2, 3) \approx 0.85712$$

Notice our estimate is only off by approximately 0.2%.

In all of the preceding random variables, it was clear what the parameters represent. For instance, in the Binomial random variable, n represents the number of trials and p represents the probability of success on any trial. And so, perhaps we are wondering what does the parameter λ represent for the Poisson random variable. Allow us to now interpret what the value of λ means.

Looking back at Interpretation 1, we see that $\lambda = np$. In the context of the previous example, we said that $\lambda = np = 100(0.02) = 2$ but allow us to now be more specific and consider the units. $\lambda = np = (100 \text{ bulbs})(2\% \text{ defective rate}) = 2 \text{ defects per box}$. Hence, we can think of λ as a *rate*.

Additionally, the following theorem which presents the expected value and variance for a Poisson random variable will give us another way to view λ .

Theorem 9.5.4

Theorem: If $X \sim \text{Poisson}(\lambda)$ then

$$\begin{aligned}\mathbb{E}[X] &= \lambda \\ \text{Var}[X] &= \lambda\end{aligned}$$

Hence, λ is not only a rate, but is also the average and variance of the Poisson distribution.

✓ Example 9.5.5

Customers arrive at a restaurant at a rate of 10 customers per hour. Assuming that the number of arrivals follows a Poisson distribution, find the probability that more than 14 customers arrive within a particular hour.

Answer

Let X model the number of customers that arrive per hour. The question tells us that the rate is 10 customers per hour and so $\lambda = 10$. Thus, $X \sim \text{Poisson}(10)$. Hence,

$$P(\text{more than 14 customers arrive within a particular hour}) = P(X > 14) = 1 - P(X \leq 14) = 1 - \text{poissoncdf}(10, 14) \approx 0.0834584728$$

Theorem 9.5.1

Theorem: Suppose there is some event which occurs according to a Poisson distribution. That is, its occurrence can be modeled by $X \sim \text{Poisson}(\lambda)$ which means the event occurs at a mean rate of λ occurrences per unit. Then the random variable Y which models the number of events occurring in the interval of length t is given by $Y \sim \text{Poisson}(\lambda t)$.

The above theorem is saying the following: suppose the number of earthquakes that occurs every year in California follows a Poisson distribution with parameter 12. That is, if X models the number of earthquakes per year in California, then $X \sim \text{Poisson}(12)$. If we were interested in modeling the number of earthquakes that occurs every two years, then this would be $\text{Poisson}(24)$. If we were interested in modeling the number of earthquakes that occurs every three years, then this would be $\text{Poisson}(36)$. If we were interested in modeling the number of earthquakes that occurs every four years, then this would be $\text{Poisson}(48)$. If we were interested in modeling the number of earthquakes that occurs every half a year, then this would be $\text{Poisson}(6)$.

With the above theorem in mind, allow us to consider the following example.

✓ Example 9.5.6

Customers arrive at a restaurant at a rate of 10 customers per hour. Assuming that the number of arrivals follows a Poisson distribution, find the probability that more than 35 customers arrive within the span of three hours.

Answer

Let X model the number of customers that arrive *per hour*. We wish to have a random variable that models the number of customers which arrive *per every three hours*. Let Y denote the random variable which models the number of customers which arrive per every three hours. By our above theorem, $Y \sim \text{Poisson}(\lambda t) = \text{Poisson}(10 \times 3) = \text{Poisson}(30)$. Hence,

$$\begin{aligned}P(\text{more than 35 customers arrive within the span of three hours}) &= P(Y > 35) = 1 - P(Y \leq 35) = 1 - \text{poissoncdf}(30, 35) \\ &\approx 0.1573834736\end{aligned}$$

We finally end this section off with an example of how conditional probabilities and random variables can be related.

✓ Example 9.5.7

The number of earthquakes in California and Alaska follows a Poisson distribution with an average of 4 earthquakes per week in California and 3 earthquakes per week in Alaska. One of these states are chosen at random and it is learned that 6 earthquakes occurred in one week. What is the probability that the state of Alaska was chosen?

Answer 1

The difficulty in this problem arises from the fact that we are unsure as to which state was picked and so allow us to introduce a partition. Let $B_1 = \{\text{California was chosen}\}$ and $B_2 = \{\text{Alaska was chosen}\}$. We have learned some event which we will denote by A so $A = \{\text{6 earthquakes occurred in one week}\}$. We are asked to find $P(B_2 | A)$ and so we have the following by Bayes' Theorem:

$$P(B_2|A) = \frac{P(B_2)P(A|B_2)}{P(B_1)P(A|B_1) + P(B_2)P(A|B_2)}$$

since both states are equally likely to be chosen

$$= \frac{0.5P(A|B_2)}{0.5P(A|B_1) + 0.5P(A|B_2)}$$

Let X and Y model the number of earthquakes in California and Alaska respectively

$$= \frac{0.5P(Y=6)}{0.5P(X=6) + 0.5P(Y=6)}$$

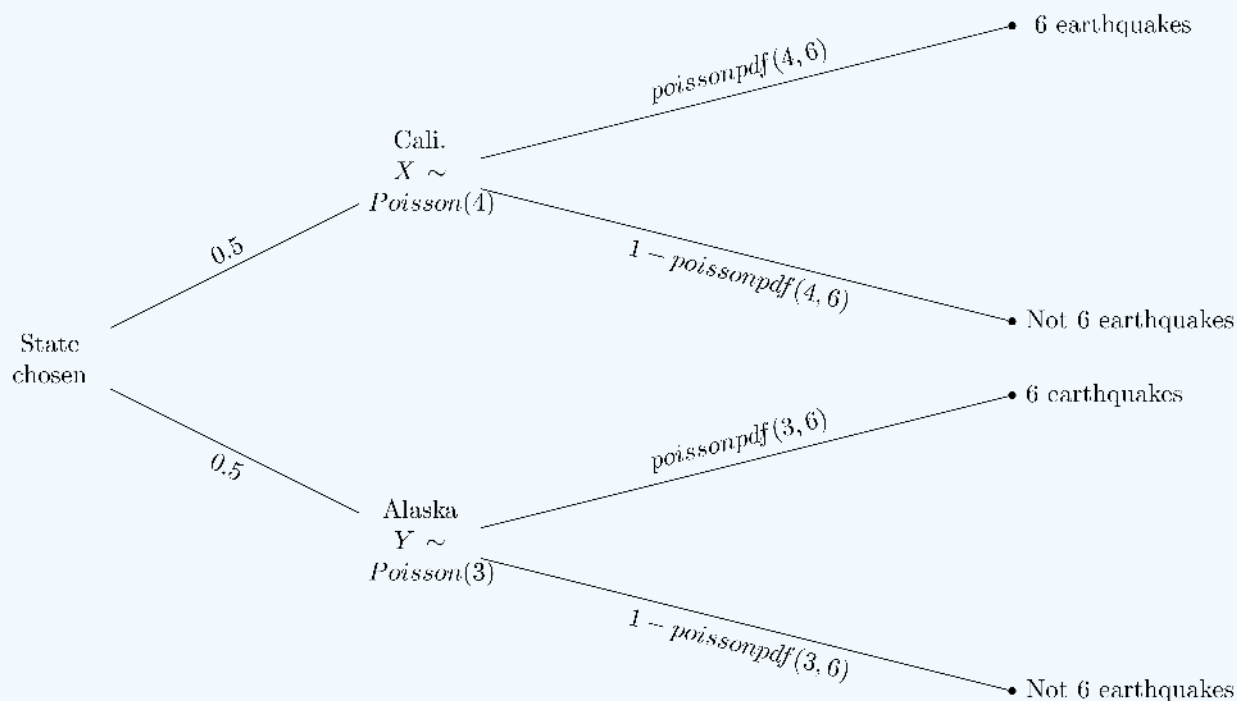
Note that $X \sim \text{Poisson}(4)$ and $Y \sim \text{Poisson}(3)$

$$= \frac{0.5\text{poissonpdf}(3,6)}{0.5\text{poissonpdf}(4,6) + 0.5\text{poissonpdf}(3,6)}$$

$$\approx 0.3260528007$$

Answer 2

Making the tree diagram for this experiment yields the following:



We are asked to find $P(\text{Alaska was chosen} | 6 \text{ earthquakes occurred})$. By Bayes' Theorem, we obtain the following:

$$P(\text{Alaska was chosen} | 6 \text{ earthquakes occurred}) = \frac{P(\text{Alaska was chosen AND 6 earthquakes occurred})}{P(6 \text{ earthquakes occurred})}$$

$$= \frac{0.5\text{poissonpdf}(3,6)}{0.5\text{poissonpdf}(4,6) + 0.5\text{poissonpdf}(3,6)}$$

$$\approx 0.3260528007$$

CHAPTER OVERVIEW

Chapter 10: Week 10

10.1: Review 2

10.2: Exam 2

Chapter 10: Week 10 is shared under a [CC BY](#) license and was authored, remixed, and/or curated by LibreTexts.

10.1: Review 2

10.1: Review 2 is shared under a [CC BY](#) license and was authored, remixed, and/or curated by LibreTexts.

10.2: Exam 2

10.2: Exam 2 is shared under a [CC BY](#) license and was authored, remixed, and/or curated by LibreTexts.

CHAPTER OVERVIEW

Chapter 11: Week 11

[11.1: Continuous Random Variables](#)

[11.2: The Uniform Continuous Distribution](#)

[11.3: The Exponential Distribution](#)

[11.4: The Normal Distribution](#)

Chapter 11: Week 11 is shared under a [CC BY](#) license and was authored, remixed, and/or curated by LibreTexts.

11.1: Continuous Random Variables

As stressed in the last chapter, the point of a random variable is to model some sort of quantity, process, or real life phenomenon. Up until now, we have discussed how discrete random variables can be used to model certain quantities; for instance, they can model the total number of successes among n independent trials or they may model the number of trials until your first success or they may model the number of occurrences of a particular event. However, discrete random variables fall short in the sense that they cannot model *every* quantity out there in the universe. Specifically, there are certain quantities which are not discrete in nature and thus will require a different type of random variable. The focus of this chapter is to discuss this new type of random variable called a *continuous random variable*. Essentially, we will have a continuous random variable whenever the quantity we wish to study or model can assume every value along some interval of real numbers. Before we rigorously define a continuous random variable, allow us to concretely understand the shortcomings of a discrete random variable and in doing so, we will motivate the need for a continuous random variable.

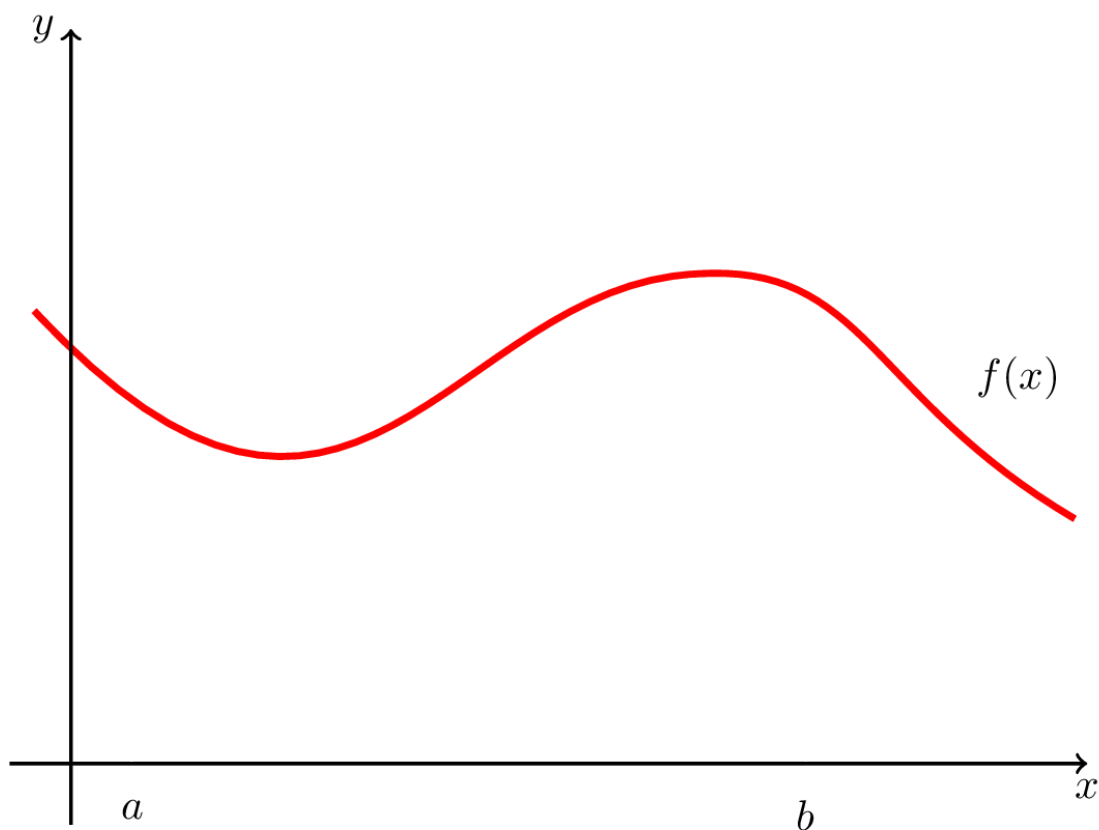
1. Suppose you wished to model the height of a person. Is height a discrete quantity or a continuous quantity? To be clear, when we say discrete quantity, we mean something along the lines of the random variable assuming the values $1, 2, 3, 4, \dots$ and whenever we say continuous, we typically mean that the random variable can assume *every value over some interval*. Clearly, height is a continuous quantity. Why? Well, if you said that height was discrete then you are (roughly) saying that a person's height can only assume values such as $1, 2, 3, 4, 5, \dots$. If we are measuring in inches, then you are saying it is impossible for someone to be $62.34254543\dots$ inches and since this may actually happen, this quantity is not discrete. Alternatively, one may think about this another way. Is it possible to grow from 5 feet 2 inches and then immediately become 5 feet 3 inches? No! When you grow, you do not grow discretely. Your body does not say we can only grow in one inch increments. Instead, as you grow from 5 feet 2 inches to 5 feet 3 inches, your height assumes every possible real number from 62 inches to 63 inches and so height must be a continuous quantity.
2. Think about the amount of rain that falls when you watch the news. Or better yet, imagine you were measuring the rain fall in your backyard. At some time, t_1 , you may have recorded one inch of rain and at a later moment in time, t_2 , you may have measured 2 inches of rain. As the rainfall increased from 1 inch to 2 inches, the amount of rain that fell assumed every real number along that interval!
3. Yet another example of something that fails to be discrete is the waiting time between events that occur according to a Poisson Process. For instance, let us suppose that the number of earthquakes that occurs follows a Poisson random variable and let us define a random variable X that models the amount of time that passes between the first and second earthquake. Is X discrete? Meaning, does the earthquake wait to occur on only integer or rational values of time? Nope! The earthquake can happen at *any* given instance and so we can think of X as being continuous on $(0, \infty)$. For more information about waiting times a good idea is to search "Queueing Theory".
4. Another example we may consider is what if X models picking a number at random from the interval $[0, 1] = \{x \in \mathbb{R} \mid 0 \leq x \leq 1\}$. For those with a background in discrete mathematics, you may recognize this to be an *uncountably infinite* set and thus X cannot be a discrete random variable since by definition, a discrete random variable takes on at most a countably infinite number of values.
5. Finally, consider the amount of miles a car gets per gallon. Since the amount of miles can assume any real number over some interval, X is a continuous random variable.

In summary, discrete random variables fail to model many quantities and thus a new class of random variables are needed. This class is called continuous random variables.

Definition: Continuous Random Variable

Definition: We say that a random variable X has a *continuous distribution* or that X is a *continuous random variable* if there exists a nonnegative function f , defined on the real line, such that for every interval of real numbers, the probability that X takes a value in the interval is the integral of f over the interval. We call this nonnegative function, f , the *probability density function* (pdf) of X .

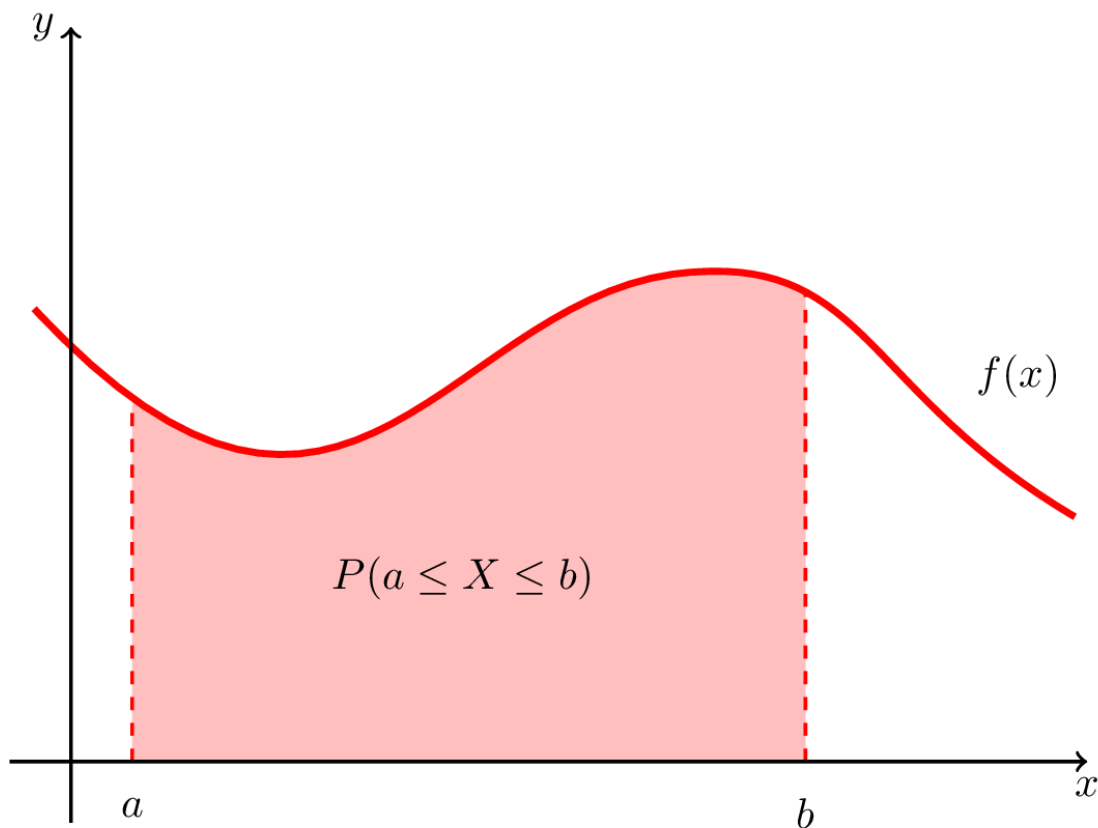
In terms of a diagram, the above definition is saying the following: suppose we have a continuous random variable X that has the following probability density function:



If we wish to find $P(a \leq X \leq b)$ then all we have to do is simply integrate f from a to b . That is, for a continuous random variable,

$$P(a \leq X \leq b) = \int_a^b f(x) dx \quad (11.1.1)$$

which geometrically represents the *area* underneath the curve of f from a to b as pictured below:



More generally, the above definition is saying

$$P(X \in I) = \int_I f(x) dx \quad (11.1.2)$$

. In words, this is saying for any continuous random variable, X to find the probability that X belongs to some interval of real numbers, we simply integrate the density function over that interval.

Believe it or not, when it comes to the theory of continuous random variables, we pretty much already know it! Before we explain why, allow us to consider a quick example based on the above definition.

✓ Example 11.1.1

Suppose a random variable X has the following density:

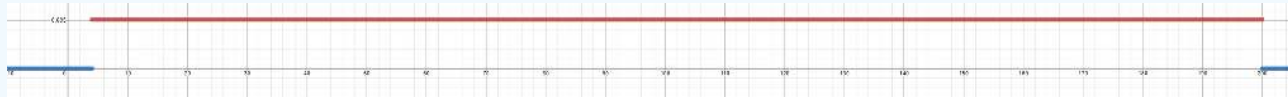
$$f(x) = \begin{cases} \frac{1}{196} & \text{if } 4 \leq x \leq 200 \\ 0 & \text{otherwise} \end{cases}$$

Find

1. $P(20 \leq X \leq 50)$
2. $P(X \leq 50) = F(50)$
3. $P(X = 20)$

Answer

1. When it comes to continuous random variables, we should always have a picture in mind and so allow us to graph the density function. Doing so yields:



Based off our definition of a continuous random variable,

$$P(X \in I) = \int_I f(x) dx$$

and so

$$P(20 \leq X \leq 50) = \int_{20}^{50} f(x) dx = \int_{20}^{50} \frac{1}{196} dx = \frac{30}{196}$$

2. Again, we must simply integrate over the appropriate region while recalling from calculus how to integrate a piecewise defined function. Hence, we obtain the following:

$$\begin{aligned} F(50) = P(X \leq 50) &= \int_{-\infty}^{50} f(x) dx \\ &= \int_{-\infty}^4 0 dx + \int_4^{50} \frac{1}{196} dx \\ &= 0 + \int_4^{50} \frac{1}{196} dx \\ &= \frac{46}{196} \end{aligned}$$

3. For this part, we again must integrate f over the appropriate region. The interval $\{X = 20\}$ is simply the single point $x = 20$ and so we have

$$P(X = 20) = \int_{20}^{20} f(x) dx = 0$$

The third part of the above problem shows us an extremely important result which is true for all continuous distributions.

Theorem: The Probability at a Single Point 11.1.1

Theorem: If X is a continuous random variable, then $P(X = a) = 0$.

Proof:

Let f denote the probability density function of X . Then $P(X = a) = \int_a^a f(x) dx = 0$.

The formal proof is quite simple and we should recall the intuitive idea of what is happening here. Essentially, to find the probability that X equals to a single point, by definition, we must integrate over that region which geometrically means find the area underneath the curve. But that region is simply a one-dimensional line and the area underneath the curve of any one-dimensional line is always 0. So again, for any continuous random variable, the probability that the random variable equals a single value is always zero. We only start to get nonzero probabilities whenever we integrate over a (non-zero) region.

As a consequence of the above discussion, please note that unlike discrete random variables, when it comes to finding probabilities for continuous random variables, it does not matter if the endpoints of an interval are included or not since the probability X equals to these values would be zero. Let us verify this claim in the next example.

Example 11.1.1

Suppose a random variable X has the following density:

$$f(x) = \begin{cases} \frac{1}{196} & \text{if } 4 \leq x \leq 200 \\ 0 & \text{otherwise} \end{cases}$$

Find $P(20 < X < 50)$.

Answer

To find $P(20 < X < 50)$ we simply integrate f over this region.

$$\begin{aligned} P(20 < X < 50) &= \int_{20}^{50} f(x) dx \\ &= \int_{20}^{50} \frac{1}{196} dx \\ &= \frac{30}{196} \end{aligned}$$

However, some students tend to object to this computation. They often feel like the limits of integration are incorrect even though they are not. Thus, allow me to argue this another way.

$$\begin{aligned}
 \{20 \leq X \leq 50\} &= \{X = 20\} \cup \{20 < X < 50\} \cup \{X = 50\} \\
 \rightarrow P(20 \leq X \leq 50) &= P(X = 20) + P(20 < X < 50) + P(X = 50) \\
 \rightarrow P(20 \leq X \leq 50) &= 0 + P(20 < X < 50) + 0 \\
 \rightarrow P(20 \leq X \leq 50) &= P(20 < X < 50) \\
 \rightarrow \frac{30}{196} &= P(20 < X < 50)
 \end{aligned}$$

Earlier we said when it comes to the theory of continuous random variables, we pretty much already know the theory. We are now in a position to say why. Informally, we make the following comment:

Note

Loosely speaking, if we replace the summation sign in the theory for discrete random variables with an integral, then we obtain the theory for continuous random variables.

More formally, we present the following chart which gives a summary of the theory required for discrete random variables and continuous random variables. In doing so, we see how similar the two types of random variables are.

Discrete Random Variables	Continuous Random Variables
$f(x)$ is called the probability mass function (pmf)	$f(x)$ is called the probability density function (pdf)
$f(x) \geq 0$ for all x	$f(x) \geq 0$ for all x
$\sum_{\text{all } x} f(x) = 1$	$\int_{\text{all } x} f(x) dx = 1$ or equivalently $\int_{-\infty}^{\infty} f(x) dx = 1$
$P(X \in A) = \sum_{\text{all } x \in A} f(x)$	$P(X \in I) = \int_I f(x) dx$
The cumulative distribution function is typically denoted by F	The cumulative distribution function is typically denoted by F
$F(x) = P(X \leq x) = \sum_{\text{all values less than or equal to } x} f(x)$	$F(x) = P(X \leq x) = \int_{\text{all values less than or equal to } x} f(t) dt = \int_{-\infty}^x f(t) dt$
$pmf \xrightarrow{\text{summing}} cdf$	$pdf \xrightarrow{\text{integrating}} cdf$
$cdf \xrightarrow{\text{subtracting}} pmf$	$cdf \xrightarrow{\text{differentiating}} pdf$
$\mathbb{E}[X] = \sum_{\text{all } x} x f(x)$	$\mathbb{E}[X] = \int_{\text{all } x} x f(x) dx = \int_{-\infty}^{\infty} x f(x) dx$
LOTUS: $\mathbb{E}[g(X)] = \sum_{\text{all } x} g(x) f(x)$	LOTUS: $\mathbb{E}[g(X)] = \int_{\text{all } x} g(x) f(x) dx = \int_{-\infty}^{\infty} g(x) f(x) dx$
$\text{Var}[X] = \mathbb{E}[X^2] - \left(\mathbb{E}[X]\right)^2$	$\text{Var}[X] = \mathbb{E}[X^2] - \left(\mathbb{E}[X]\right)^2$
$SD[X] = \sqrt{\text{Var}[X]}$ (again we only care for the positive square root)	$SD[X] = \sqrt{\text{Var}[X]}$ (again we only care for the positive square root)

Caution

Notice a fundamental difference between the probability mass function and the probability density function. In the probability mass function, we obtain probabilities by *summing over the appropriate values of x when plugged into the function, f* . However, with a probability density function, we obtain probabilities by *integrating the function over the appropriate region*. Thus we should NEVER plug a value into a density function and think that the value obtained is the corresponding probability. For instance, with regards to Example 1, the following would be incorrect: $P(X = 20) = f(20) = \frac{1}{196}$. We are only allowed to plug in if we have a discrete random variable!

In short, plugging in values into a density does not yield a probability. Instead, we only obtain a probability once we integrate the density over the appropriate interval.

Once you know the above chart, you can pretty much work with any continuous random variable. Let us touch upon every aspect of that chart by considering the following examples.

✓ Example 11.1.1

Suppose X has the following density: $f(x) = \begin{cases} cx^2 & \text{if } 2 \leq x \leq 4 \\ 0 & \text{otherwise} \end{cases}$

- 1) Find the value of c .
- 2) $P(2 \leq X \leq 3)$.
- 3) Is $P(-10 \leq X \leq 3) = P(2 \leq X \leq 3)$?
- 4) Find the cumulative distribution function of X .
- 5) Using the cumulative distribution function of X find $P(2 \leq X \leq 3)$.
- 6) Find $\mathbb{E}[X]$.
- 7) Find $\text{Var}[X]$.

Answer

1) We will first walk through the full details of this problem and then at the end we will simplify it. If this was a discrete random variable, how would we find the value of c ? Think back to the homework or our review sheet. To find the value of c , we would impose the condition that $\sum_{\text{all } x} f(x) = 1$. According to our chart, what is the analogous property for the continuous case? The analogous property is that $\int_{-\infty}^{\infty} f(x) dx = 1$. We impose this condition and we may easily find c .

$$\int_{-\infty}^{\infty} f(x) dx = 1$$

Recall from calculus that since f is a piecewise defined function, we have to break up the integral over each piece:

$$\int_{-\infty}^{\infty} f(x) dx = 1$$

$$\int_{-\infty}^2 0 dx + \int_2^4 cx^2 dx + \int_4^{\infty} 0 dx = 1$$

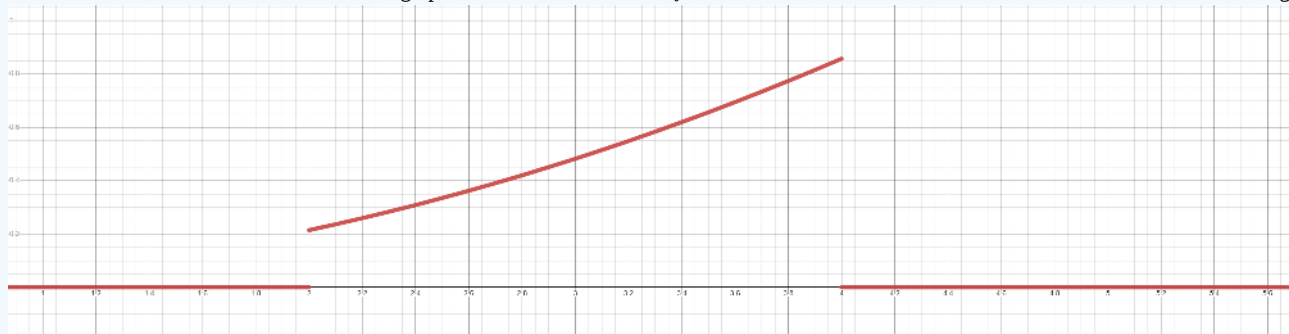
We see that the first and third integrals are simply zero. Thus we have the following calculation:

$$\begin{aligned} \int_{-\infty}^{\infty} f(x) dx &= 1 \\ \int_{-\infty}^2 0 dx + \int_2^4 cx^2 dx + \int_4^{\infty} 0 dx &= 1 \\ \int_2^4 cx^2 dx &= 1 \\ c \int_2^4 x^2 dx &= 1 \\ c \left[\frac{1}{3} x^3 \right]_2^4 &= 1 \\ c \left[\frac{64}{3} - \frac{8}{3} \right] &= 1 \\ c \left[\frac{56}{3} \right] &= 1 \\ c &= \frac{3}{56} \end{aligned}$$

Hence the probability density function is given by

$$f(x) = \begin{cases} \frac{3}{56} x^2 & \text{if } 2 \leq x \leq 4 \\ 0 & \text{otherwise} \end{cases}$$

And so the graph of f looks like the following:



Before we move onto the next part, allow us to make a quick remark which will help us in the future. We said to find the value of c we must integrate over all x our function $f(x)$. When we say *over all x* , we mean we must integrate from $-\infty$ to ∞ . However, recall what we also said earlier: *we only start to get nonzero probabilities whenever we integrate over a non-zero region*. Thus, to find the value of c we do not have to really integrate from $-\infty$ to ∞ . Instead, we only have to integrate over the values of x which has a nonzero density! Thus, to start the problem, I could begin by writing

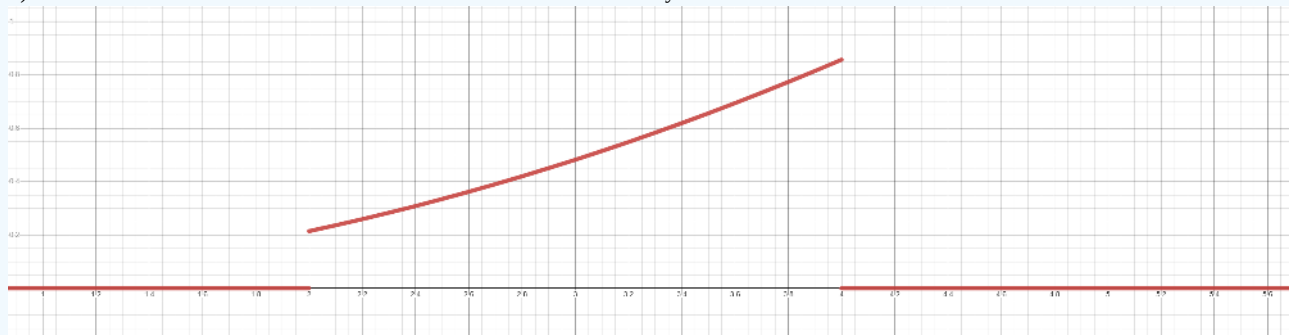
$$\int_2^4 cx^2 dx = 1$$

and proceed from there.

2) Here, we are asked to find $P(2 \leq X \leq 3)$. To find the desired probability, we simply integrate the density over the appropriate region.

$$\begin{aligned} P(2 \leq X \leq 3) &= \int_2^3 \frac{3}{56} x^2 dx \\ &= \frac{3}{56} \int_2^3 x^2 dx \\ &= \frac{3}{56} \left[\frac{1}{3} x^3 \right]_2^3 \\ &= \frac{1}{56} \left[x^3 \right]_2^3 \\ &= \frac{1}{56} [27 - 8] \\ &= \frac{19}{56} \end{aligned}$$

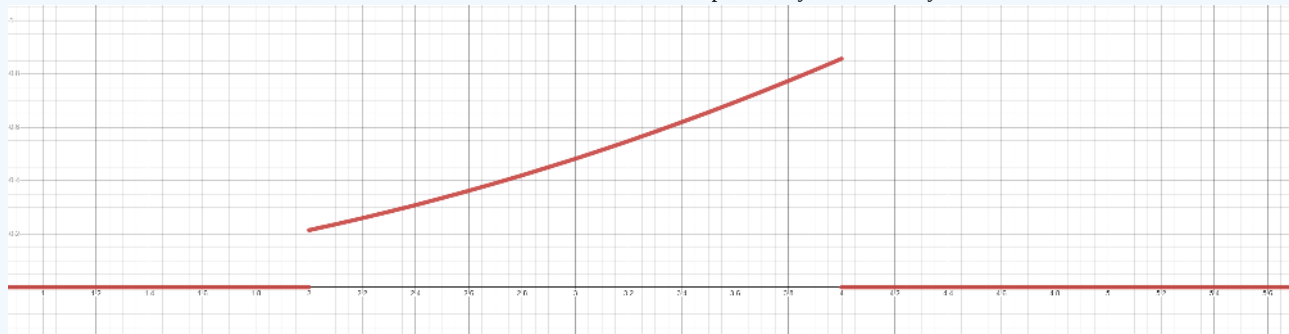
3) Recall the density looks like this:



From the picture, we see in finding $P(-10 \leq X \leq 3)$, the interval from $[-10, 2]$ does not contribute any probability. Thus, $P(-10 \leq X \leq 3)$ is indeed the same as $P(2 \leq X \leq 3)$.

4) Before we find the cumulative distribution function of X , allow us to discuss the job of a cdf for a continuous random variable. We can think of the cdf of a continuous random variable as the "area so far" function. That is, given some density f , the cdf is the function F , such that when we plug in some value of x into F , F will tell us how much area is underneath the density from the left to the value of x . \\\

As a reminder, here is what the probability density function looks like.



Since the function has three pieces, our cumulative distribution function will also have three pieces. One piece will tell us the "area so far" when $x < 2$. The other piece will tell us the "area so far" when $2 \leq x \leq 4$. The last piece will tell us the "area so far" when $x > 4$. That is, we will have the following:

$$F(x) = \begin{cases} \text{????} & \text{if } x < 2 \\ \text{????} & \text{if } 2 \leq x \leq 4 \\ \text{????} & \text{if } x > 4 \end{cases}$$

Two of these pieces are easy to find. Notice when $x < 2$ what is $F(x) = P(X \leq x)$? Well when $x < 2$, to find $F(x) = P(X \leq x)$ we have to integrate from $-\infty$ to x . Clearly this integral will equal 0 since we have "no area" from $-\infty$ to any $x < 2$.

Additionally, when $x > 4$ what is $F(x) = P(X \leq x)$? Well when $x > 4$, to find $F(x) = P(X \leq x)$ we have to integrate from $-\infty$ to x . Clearly this integral will equal 1 since we have covered the entire nonzero portion of the density. Moreover, in part 1, we said

$$\int_2^4 f(x) dx = 1$$

At this point, we are almost done. We know the following:

$$F(x) = \begin{cases} 0 & \text{if } x < 2 \\ \text{????} & \text{if } 2 \leq x \leq 4 \\ 1 & \text{if } x > 4 \end{cases}$$

All that remains is to investigate what happens to the cdf, $F(x)$ when $2 \leq x \leq 4$. To do this, we apply the rule from our chart:

$$F(x) = \int_{-\infty}^x f(t) dt$$

Before we apply this rule, allow us to be reminded of what we are searching for here. We are asking if x is some value between 2 and 4, what is the area underneath the function up until x ?

We see in the formula,

$$F(x) = \int_{-\infty}^x f(t) dt$$

We need $f(t)$. Since

$$f(x) = \begin{cases} \frac{3}{56}x^2 & \text{if } 2 \leq x \leq 4 \\ 0 & \text{otherwise} \end{cases}$$

then

$$f(t) = \begin{cases} \frac{3}{56}t^2 & \text{if } 2 \leq t \leq 4 \\ 0 & \text{otherwise} \end{cases}$$

This is just our density with the variable x being replaced by t . We now apply the formula in our chart and recall that we only care about the regions which have a non-zero density:

$$\begin{aligned}
 F(x) &= \int_{-\infty}^x f(t) dt \\
 &= \int_2^x f(t) dt \\
 &= \int_2^x \frac{3}{56} t^2 dt \\
 &= \frac{3}{56} \int_2^x t^2 dt \\
 &= \frac{3}{56} \left[\frac{1}{3} t^3 \right]_2^x \\
 &= \frac{1}{56} \left[t^3 \right]_2^x \\
 &= \frac{1}{56} [x^3 - 8] \\
 &= \frac{x^3 - 8}{56} \\
 &= \frac{1}{56} (x^3 - 8)
 \end{aligned}$$

Putting these three pieces together yields the following:

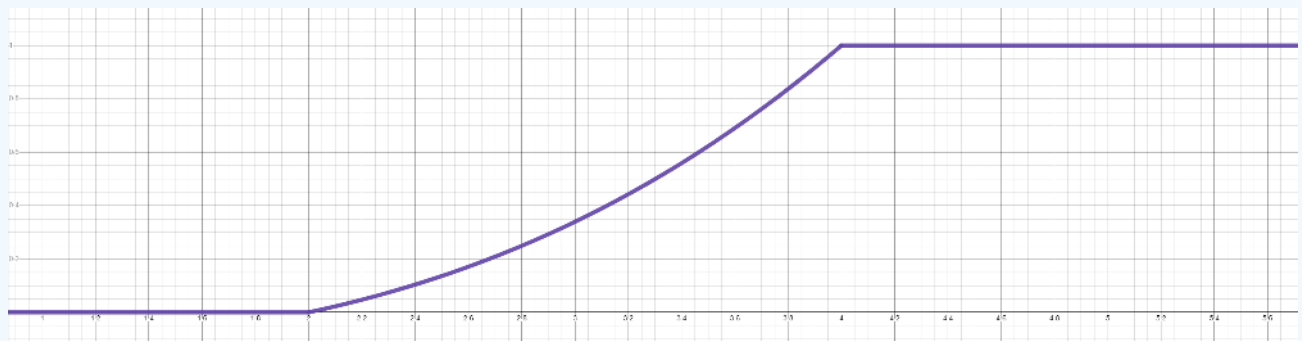
$$F(x) = \begin{cases} 0 & \text{if } x < 2 \\ \frac{1}{56}(x^3 - 8) & \text{if } 2 \leq x \leq 4 \\ 1 & \text{if } x > 4 \end{cases}$$

There are two important observations we should make note of. First, notice when we differentiate the cumulative distribution function, we get back the probability density function! Secondly, notice that the equality at each endpoint does not matter. That is, you could write your answer as

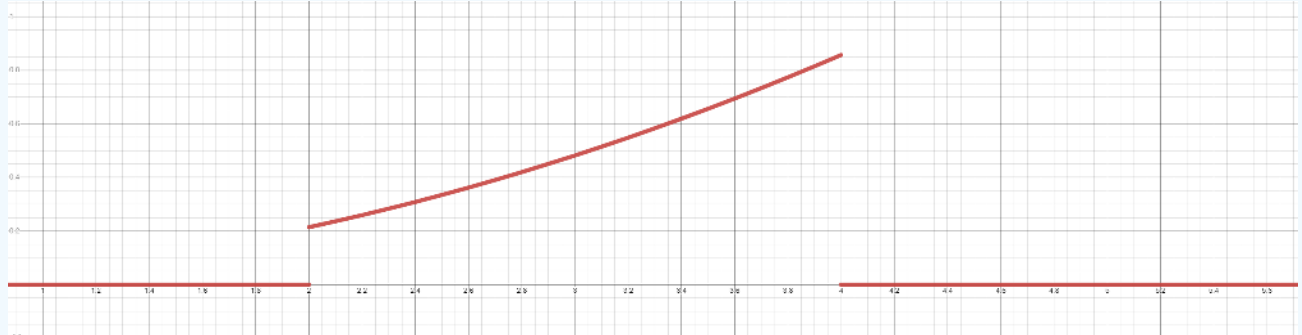
$$F(x) = \begin{cases} 0 & \text{if } x \leq 2 \\ \frac{1}{56}(x^3 - 8) & \text{if } 2 < x < 4 \\ 1 & \text{if } x \geq 4 \end{cases}$$

and this would still be correct. This stems from the fact that the probability at any single point is zero.

Although this was not asked, here is what the cumulative distribution function looks like if you were to graph it:



5) Again, consider the graph of the density function:



We wish to find $P(2 \leq X \leq 3)$. Notice that the area underneath f from 2 to 3 is precisely the area up to 3 minus the area up to 2. Hence $P(2 \leq X \leq 3) = F(3) - F(2) = \frac{19}{56} - 0 = \frac{19}{56}$.

6)

$$\begin{aligned}\mathbb{E}[X] &= \int_{-\infty}^{\infty} x f(x) dx \\ &= \int_2^4 x f(x) dx \\ &= \int_2^4 x \times \frac{3}{56} x^2 dx \\ &= \int_2^4 \frac{3}{56} x^3 dx \\ &= \frac{45}{14}\end{aligned}$$

7) To find $\mathbb{V}ar[X]$, we use the formula $\mathbb{V}ar[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$. We already found $\mathbb{E}[X]$ above. Let us use The Law of the Unconscious Statistician to find $\mathbb{E}[X^2]$.

$$\begin{aligned}\mathbb{E}[X^2] &= \int_{-\infty}^{\infty} x^2 f(x) dx \\ &= \int_2^4 x^2 f(x) dx \\ &= \int_2^4 x^2 \times \frac{3}{56} x^2 dx \\ &= \int_2^4 \frac{3}{56} x^4 dx \\ &= \frac{372}{35}\end{aligned}$$

Putting everything together, we obtain

$$\begin{aligned}\mathbb{V}ar[X] &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \\ &= \frac{372}{35} - \left(\frac{45}{14}\right)^2 \\ &= \frac{291}{980}\end{aligned}$$

✓ Example 11.1.1

Suppose the random variable X has the following probability density function:

$$f(x) = \begin{cases} ce^{-10x} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

1. Find the value of c .
2. Find $P(0 \leq X \leq 1)$.
3. Find the cumulative distribution function of X .
4. Find $\mathbb{E}[X]$.

Answer

1. Similar to the last problem, we impose the condition that $\int_{-\infty}^{\infty} f(x) dx = 1$ in order to find c .

$$\begin{aligned}
 \int_{-\infty}^{\infty} f(x) dx &= 1 \\
 \int_{-\infty}^0 0 dx + \int_0^{\infty} ce^{-10x} dx &= 1 \\
 \int_0^{\infty} ce^{-10x} dx &= 1 \\
 c \int_0^{\infty} e^{-10x} dx &= 1 \\
 c \lim_{t \rightarrow \infty} \left[\int_0^t e^{-10x} dx \right] &= 1 \\
 c \lim_{t \rightarrow \infty} \left[-\frac{1}{10} e^{-10x} \Big|_0^t \right] &= 1 \\
 -\frac{c}{10} \lim_{t \rightarrow \infty} \left[e^{-10x} \Big|_0^t \right] &= 1 \\
 -\frac{c}{10} \lim_{t \rightarrow \infty} \left[e^{-10t} - e^{-10(0)} \right] &= 1 \\
 -\frac{c}{10} \lim_{t \rightarrow \infty} \left[e^{-10t} - 1 \right] &= 1 \\
 -\frac{c}{10} \lim_{t \rightarrow \infty} \left[\frac{1}{e^{10t}} - 1 \right] &= 1 \\
 -\frac{c}{10} \left[0 - 1 \right] &= 1 \\
 \frac{c}{10} &= 1 \\
 c &= 10
 \end{aligned}$$

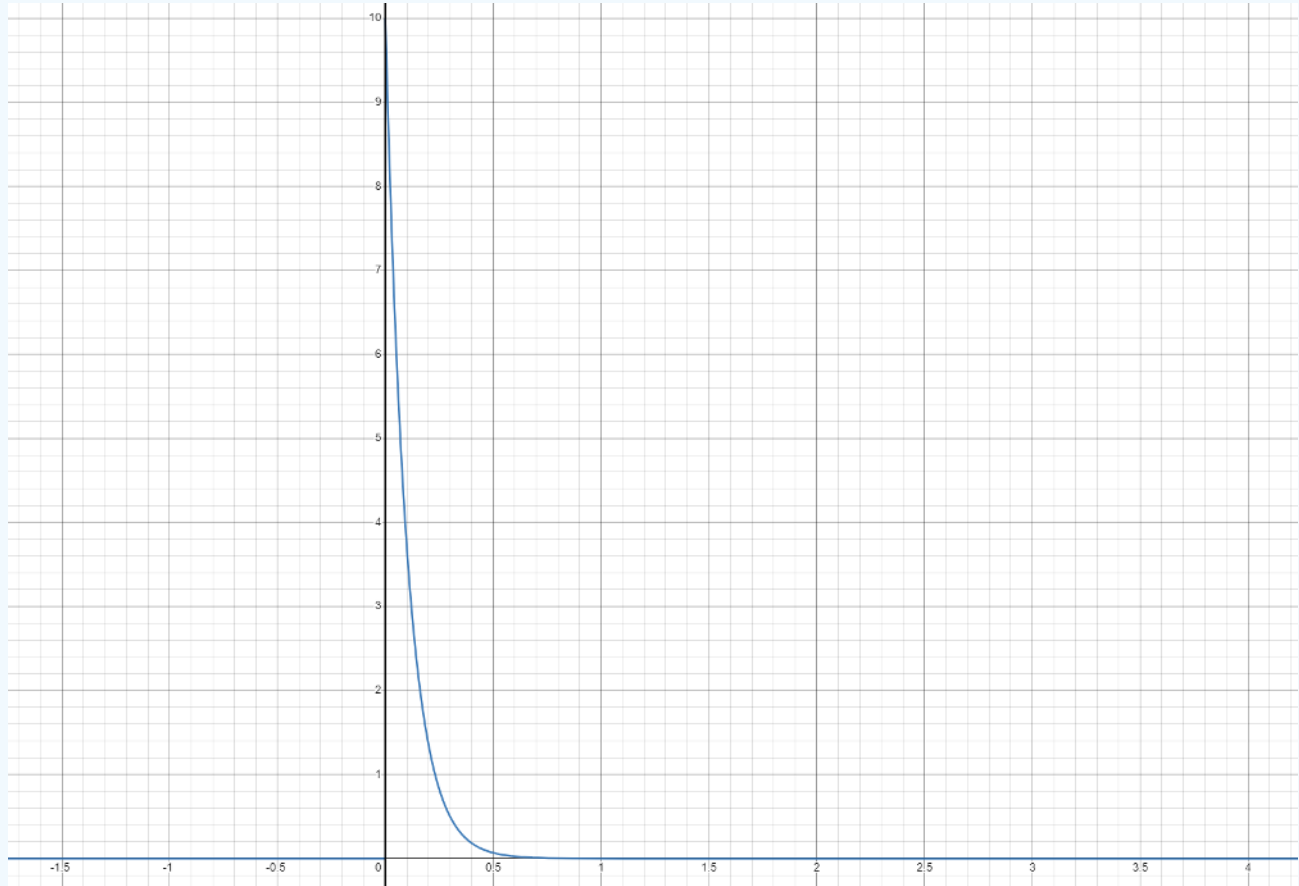
Hence the probability density function is given by

$$f(x) = \begin{cases} 10e^{-10x} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

2. Here, we are asked to find $P(0 \leq X \leq 1)$. To find the desired probability, we simply integrate the density over the appropriate region.

$$\begin{aligned}
 P(0 \leq X \leq 1) &= \int_0^1 f(x) dx \\
 &= \int_0^1 10e^{-10x} dx \\
 &= -e^{-10x} \Big|_0^1 \\
 &= -e^{-10} - (-e^0) \\
 &= -e^{-10} - (-1) \\
 &= 1 - e^{-10}
 \end{aligned}$$

3. Here is what the density function looks like:



Clearly, when $x \leq 0$, $F(x) = 0$ since we have no area from $-\infty$ to any $x \leq 0$.

So we have the following piecewise defined function:

$$F(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ \text{????} & \text{if } x > 0 \end{cases}$$

To find out what the cdf is when $x > 0$ we use the rule from our chart. When $x > 0$ we have the following:

$$\begin{aligned} F(x) &= \int_{-\infty}^x f(t) dt \\ &= \int_0^x f(t) dt \\ &= \int_0^x 10e^{-10t} dt \\ &= -e^{-10t} \Big|_0^x \\ &= -e^{-10x} - (-e^{-10(0)}) \\ &= -e^{-10x} - (-1) \\ &= 1 - e^{-10x} \end{aligned}$$

And so we have

$$F(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ 1 - e^{-10x} & \text{if } x > 0 \end{cases}$$

Notice that as $x \rightarrow \infty$, $F(x) \rightarrow 1$ and also observe that when we differentiate the cumulative distribution function, we get back the probability density function.

4.

$$\begin{aligned}
 \mathbb{E}[X] &= \int_{-\infty}^{\infty} x f(x) dx \\
 &= \int_{-\infty}^0 x f(x) dx + \int_0^{\infty} x f(x) dx \\
 &= \int_{-\infty}^0 x(0) dx + \int_0^{\infty} x(10e^{-10x}) dx \\
 &= 0 + \int_0^{\infty} x(10e^{-10x}) dx \\
 &= \int_0^{\infty} x(10e^{-10x}) dx && \text{Improper Integral} \\
 &= \lim_{t \rightarrow \infty} \left[\int_0^t x \times 10e^{-10x} dx \right] && \text{Integration by parts with } u = x, dv = 10e^{-10x} dx \\
 &= \lim_{t \rightarrow \infty} \left[-xe^{-10x} \Big|_0^t - \int_0^t -e^{-10x} dx \right] \\
 &= \lim_{t \rightarrow \infty} \left[(-te^{-10t} + 0) + \int_0^t e^{-10x} dx \right] \\
 &= \lim_{t \rightarrow \infty} \left[-te^{-10t} + \left(-\frac{1}{10}e^{-10x} \right) \Big|_0^t \right] \\
 &= \lim_{t \rightarrow \infty} \left[-\frac{t}{e^{10t}} + \left(-\frac{1}{10}e^{-10t} + \frac{1}{10} \right) \right] \\
 &= \lim_{t \rightarrow \infty} \left[-\frac{t}{e^{10t}} - \frac{1}{10}e^{-10t} + \frac{1}{10} \right] && \text{L'Hospital's Rule} \\
 &= \frac{1}{10}
 \end{aligned}$$

11.2: The Uniform Continuous Distribution

As mentioned in Week 9, there are certain processes which occurs more frequently than others and so we will attach "brand" names to the random variables which models these processes. In terms of discrete random variables, we studied the (discrete) Uniform Random Variable, The Bernoulli and Binomial Random Variables, The Geometric and Negative Binomial Random Variables, and the Poisson Random Variable.

For the class of continuous random variables, we will study the (continuous) Uniform Random Variable, the Exponential Random Variable, and the Normal Random Variable. We first begin with the (continuous) Uniform Random Variable.

Definition: The (Continuous) Uniform Random Variable

Definition: Let $a < b$ be real numbers. Suppose X is a random variable who has the following density:

$$f(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a \leq x \leq b \\ 0, & \text{otherwise} \end{cases} \quad (11.2.1)$$

Then we say X has the *uniform distribution* on the interval $[a, b]$ and we write $X \sim \mathcal{U}[a, b]$.

Notice that the (continuous) Uniform Random Variable is something which is not totally new for us. Observe that we actually studied the $\mathcal{U}[4, 200]$ random variable in the last section and so many of the proofs are left to the reader. Additionally, recall that the point of a random variable is to model some sort of quantity under certain assumptions and so we can ask, "what does the (continuous) Uniform random variable model?"

Interpretation of the (Discrete) Uniform Random Variable

The (continuous) Uniform Random Variable models the outcome of an experiment where the experiment is choosing a number at random from the interval $[a, b]$.

We end this section with a theorem concerning some of the properties of a discrete Uniform Random Variable. Their proofs are left as homework exercises.

Theorem 11.2.1

Theorem: If $X \sim \mathcal{U}[a, b]$ then:

$$\begin{aligned} \mathbb{E}[X] &= \frac{a+b}{2} \\ \mathbb{V}ar[X] &= \frac{(b-a)^2}{12} \\ F(x) &= \begin{cases} 0, & \text{if } x < a \\ \frac{x-a}{b-a}, & \text{if } a \leq x < b \\ 1, & \text{if } x \geq b \end{cases} \end{aligned}$$

11.2: The Uniform Continuous Distribution is shared under a [CC BY](#) license and was authored, remixed, and/or curated by LibreTexts.

11.3: The Exponential Distribution

Here, we will discuss another brand name random variable, the Exponential random variable which is related to the Poisson random variable discussed a few sections ago.

Definition: The Exponential Random Variable

Definition: Let $\theta > 0$. Suppose X is a random variable who has the following density function:

$$f(x) = \begin{cases} \frac{1}{\theta} e^{-x/\theta} & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

Then we say X is an *Exponential random variable* with parameter θ and we write $X \sim \text{Exp}(\theta)$.

Notice that the Exponential random variable is something which is not totally new for us since we actually studied the $\text{Exp}(\frac{1}{10})$ random variable in the last section! Additionally, recall that the point of a random variable is to model some sort of quantity under certain assumptions and so we can ask, "what does the Exponential random variable model?"

Interpretation of the Exponential Random Variable

Recall that the Poisson random variable models the number of occurrences in a given interval (for instance, the number of earthquakes per week). This number was a discrete-type of random variable that follows the Poisson distribution with parameter or rate or mean λ . However, suppose we were interested in modeling the *waiting time* between successive occurrences (for instance, we wished to model the time elapsed between two successive earthquakes). Then it can be shown that the waiting time between the next occurrence is an Exponential random variable with parameter $\theta = \frac{1}{\lambda}$.

Theorem 11.3.1

Theorem: If $X \sim \text{Exp}(\theta)$ then

$$\mathbb{E}[X] = \theta \text{ and } \text{Var}[X] = \theta^2$$

Theorem 11.3.2

Theorem: If $X \sim \text{Exp}(\theta)$ then the cumulative distribution of X is given by:

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - e^{-x/\theta}, & \text{if } x \geq 0 \end{cases}$$

Example 11.3.3

Customer arrive in a certain shop according to a Poisson process with a mean rate of $\lambda = 1/3$ customers per minute. What is the probability that the shopkeeper will have to wait more than five minutes for the arrival of the first customer?

Answer

According to the interpretation of the Exponential random variable, if we have a Poisson process with rate λ , then the waiting time for the next customer is going to be an exponential random variable with parameter $\theta = \frac{1}{\lambda}$. Since customers arrive to the store with rate $\lambda = 1/3$, then the waiting time for the next customer is going to be distributed as an Exponential random variable with parameter $\theta = \frac{1}{\lambda} = \frac{1}{1/3} = 3$. So if X models the waiting time for the next customer, then it follows that $X \sim \text{Exp}(3)$. That is, X has the following density:

$$f(x) = \begin{cases} \frac{1}{3} e^{-x/3} & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

Hence

$$\begin{aligned}
 P(\text{the waiting time is more than 5 minutes}) &= P(X > 5) \\
 &= \int_5^{\infty} \frac{1}{3} e^{-x/3} dx \\
 &= e^{-5/3} \\
 &\approx 0.1889
 \end{aligned}$$

Note that we can achieve the same answer by utilizing the cdf:

$$\begin{aligned}
 P(\text{the waiting time is more than 5 minutes}) &= P(X > 5) \\
 &= 1 - P(X \leq 5) \\
 &= 1 - F(5) \\
 &= 1 - (1 - e^{-5/3}) \\
 &= e^{-5/3} \\
 &\approx 0.1889
 \end{aligned}$$

✓ Example 11.3.4

Example: Suppose that a certain type of electronic component has an exponential distribution with a mean life of 500 hours. If X denotes the lifetime of this component then find $P(X > 900 | X > 300)$.

Answer 1

Let X model the lifetime of the component. From the problem, we know that $X \sim \text{Exp}(\theta)$. The problem tells us the mean is 500 hours and for an Exponential random variable, the mean and parameter are the same thing! So we know that $X \sim \text{Exp}(500)$. That is, the density is given by

$$f(x) = \begin{cases} \frac{1}{500} e^{-x/500} & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

Thus we have

$$\begin{aligned}
 P(X > 900 | X > 300) &= \frac{P(\{X > 900\} \cap \{X > 300\})}{P(X > 300)} \\
 &= \frac{P(X > 900)}{P(X > 300)} \\
 &= \frac{\int_{900}^{\infty} \frac{1}{500} e^{-x/500} dx}{\int_{300}^{\infty} \frac{1}{500} e^{-x/500} dx} \\
 &= \frac{e^{-900/500}}{e^{-300/500}} \\
 &= e^{-6/5} \\
 &\approx 0.3012
 \end{aligned}$$

Answer 2

Or, as noted in the last problem, we can use the cdf to our advantage to avoid the integration.

Alternative Solution:

$$\begin{aligned}P(X > 900 | X > 300) &= \frac{P(\{X > 900\} \cap \{X > 300\})}{P(X > 300)} \\&= \frac{P(X > 900)}{P(X > 300)} \\&= \frac{1 - P(X \leq 900)}{1 - P(X \leq 300)} \\&= \frac{1 - F(900)}{1 - F(300)} \\&= \frac{1 - (1 - e^{-900/500})}{1 - (1 - e^{-300/500})} \\&= \frac{e^{-900/500}}{e^{-300/500}} \\&= e^{-6/5} \\&\approx 0.3012\end{aligned}$$

11.3: The Exponential Distribution is shared under a [CC BY](#) license and was authored, remixed, and/or curated by LibreTexts.

11.4: The Normal Distribution

In this section, we are going to study what is arguably the most important continuous random variable - the normal random variable. As usual, our process is to first define the density of this random variable.

Definition: The Normal Random Variable

Definition: Let μ and σ be numbers satisfying $-\infty < \mu < \infty$ and $\sigma > 0$ (so μ can be any real number and σ is any positive real number). Suppose X is a random variable who has the following density function:

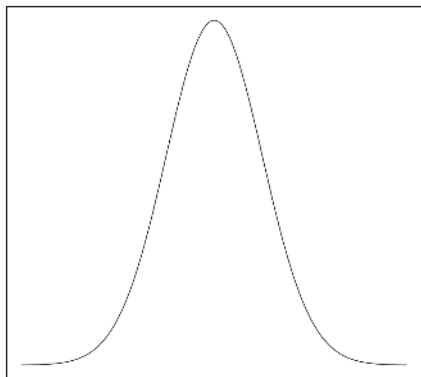
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad \text{if } -\infty < x < \infty \quad (11.4.1)$$

or equivalently

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \quad \text{if } -\infty < x < \infty \quad (11.4.2)$$

Then we say X is a **Normal random variable** with parameters μ and σ^2 we write $X \sim \mathcal{N}(\mu, \sigma^2)$

We will see why the Normal distribution is important in the next section. For now, allow us to discuss the properties of this distribution. Although the density presented in our definition looks extremely complicated, the graph of the density is something you may have seen before:

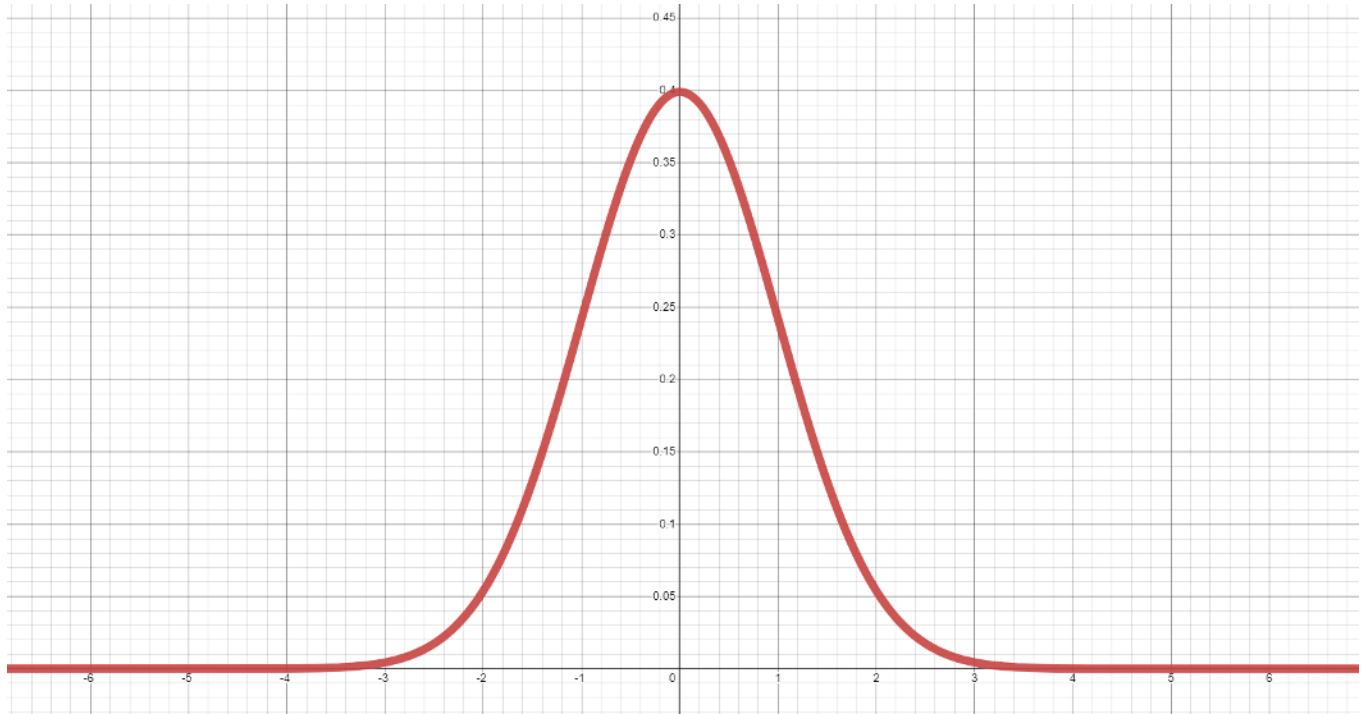


Specifically, the graph of the density has what we refer to as a bell-shaped curve. As with all the other random variables, let us take a moment to consider how each parameter influences the density. Essentially, the first parameter will tell us the "center" of the distribution or the x -value for which the distribution is symmetric about. Meanwhile, the second parameter tells us how "thick/thin" the distribution is. In order to see this, allow us to graph the density of a few normal random variables. We will first change the value of μ while keeping σ^2 the same.

Allow us to graph the $\mathcal{N}(0, 1)$ random variable which has density

$$\frac{1}{\sqrt{1}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-0}{\sqrt{1}}\right)^2} = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

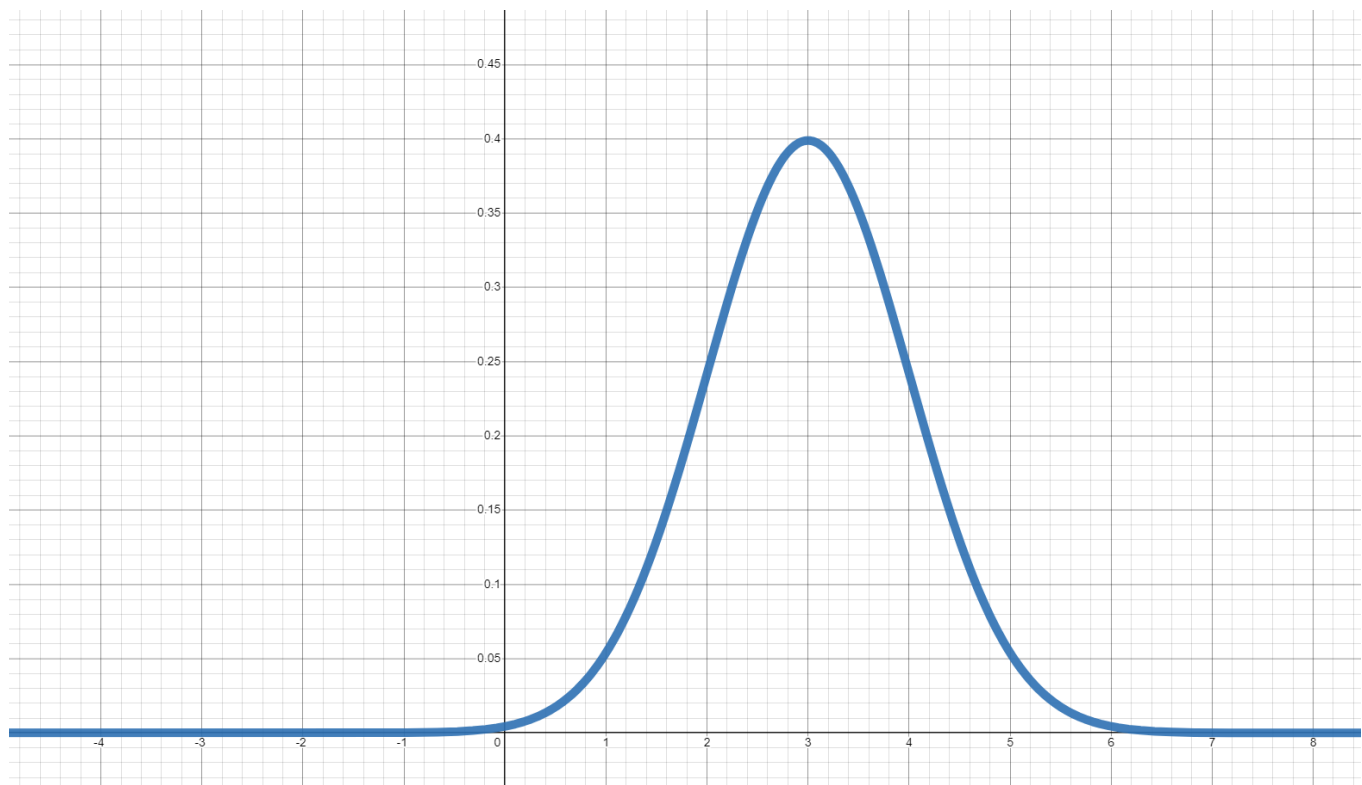
Graphing this function produces the following curve:



Notice for the $\mathcal{N}(0, 1)$ random variable, the center of the distribution is at $x = 0$. Let us now change the first parameter to see how this affects the graph. Sketching the density of the $\mathcal{N}(3, 1)$ random variable which has density

$$\frac{1}{\sqrt{1}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-3}{\sqrt{1}}\right)^2} = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-3)^2}{2}}$$

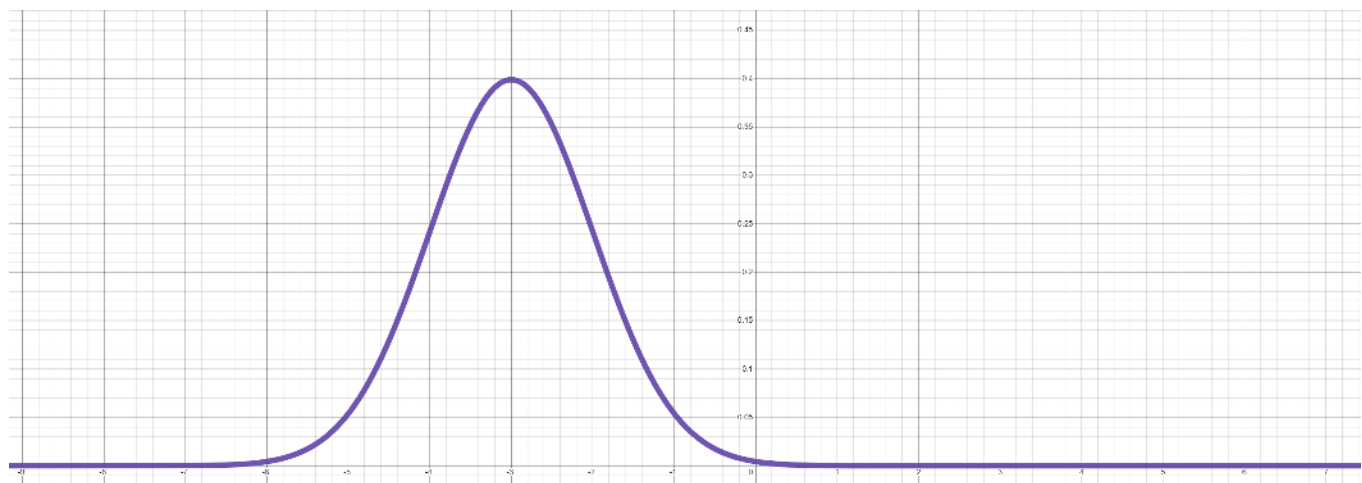
yields:



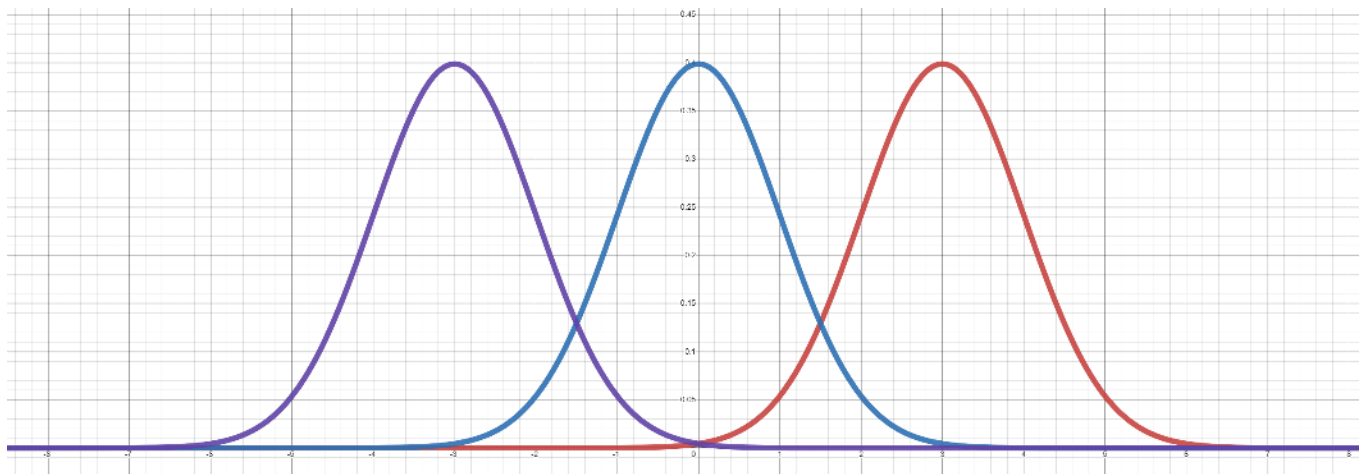
Notice for the $\mathcal{N}(3, 1)$ random variable, the center of the distribution is at $x = 3$. And finally, sketch the density of the $\mathcal{N}(-3, 1)$ random variable which has density

$$\frac{1}{\sqrt{1}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x+3}{\sqrt{1}}\right)^2} = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x+3)^2}{2}}$$

yields:

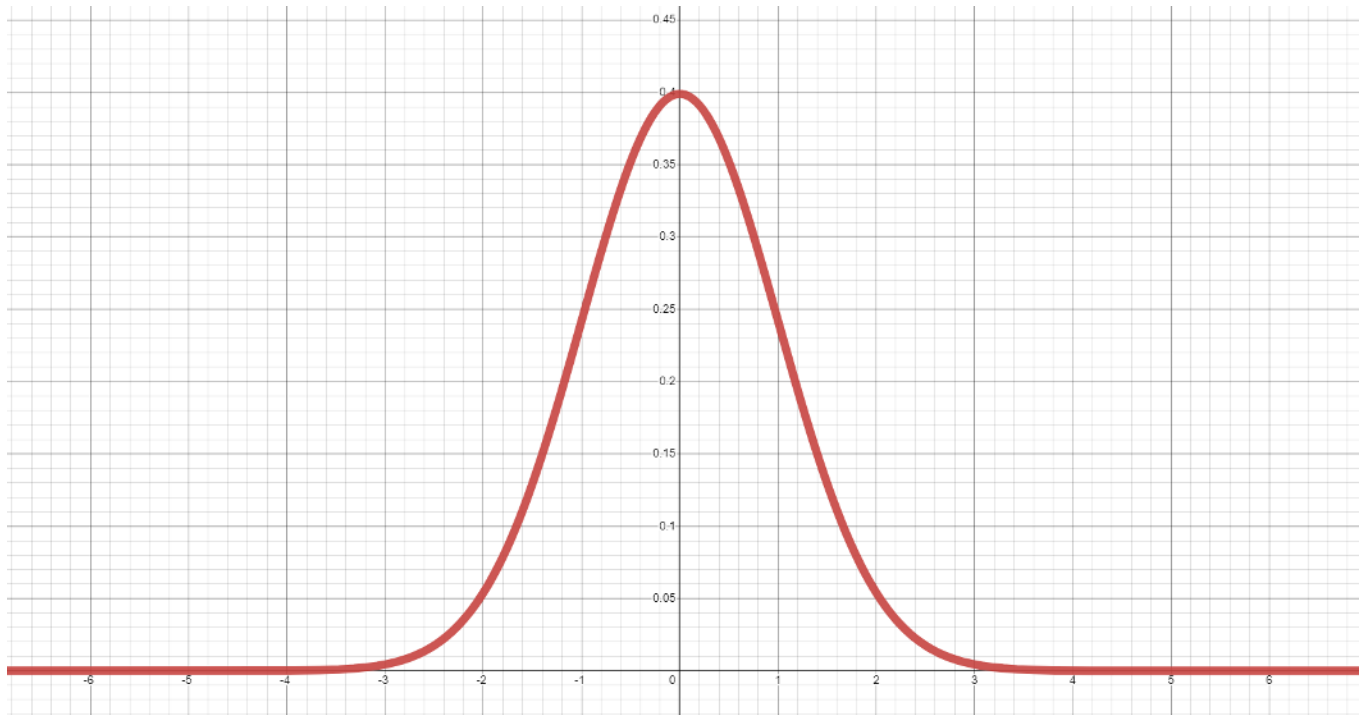


Notice for the $\mathcal{N}(-3, 1)$ random variable, the center of the distribution is at $x = -3$. Placing all three densities side by side yields the following:

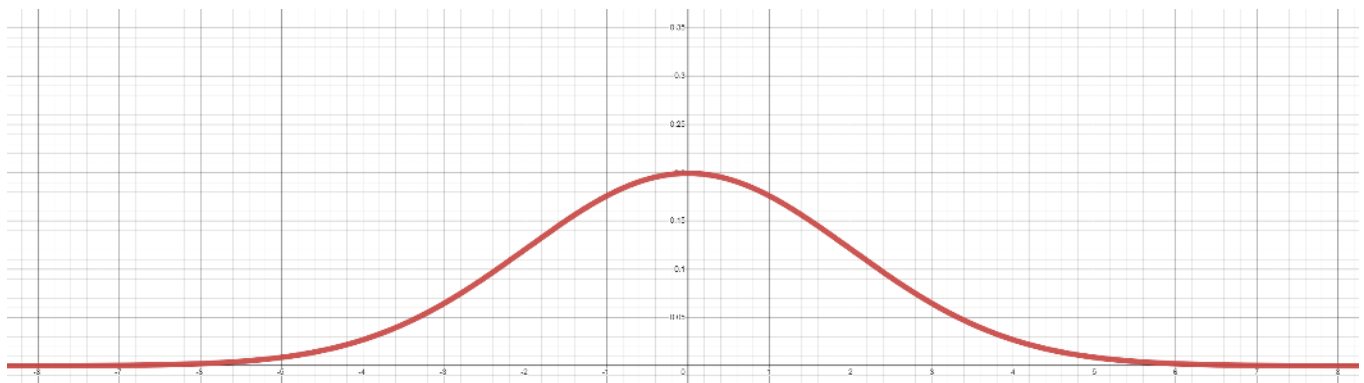


Intuitively, we see the first parameter controls the center of the distribution.

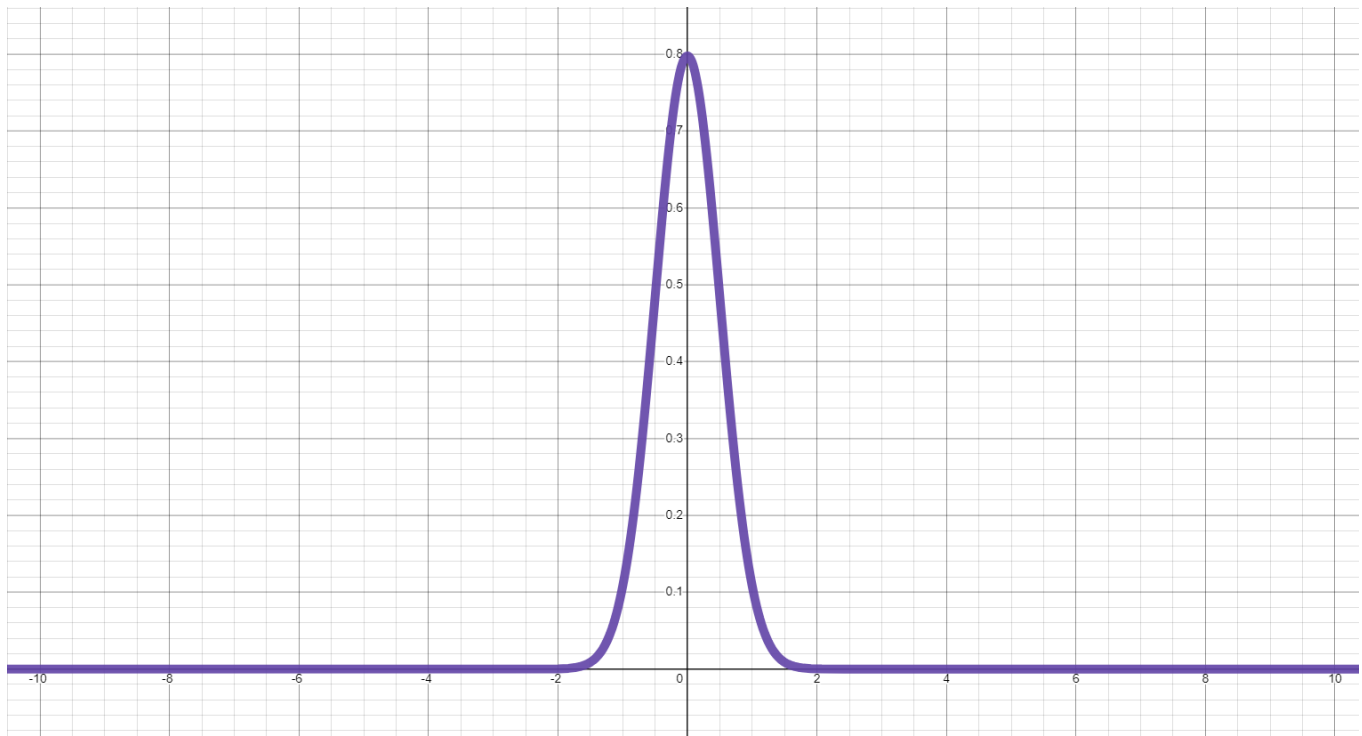
We now perform a similar process to the second parameter. Let us again start with the density of the $\mathcal{N}(0, 1)$ random variable:



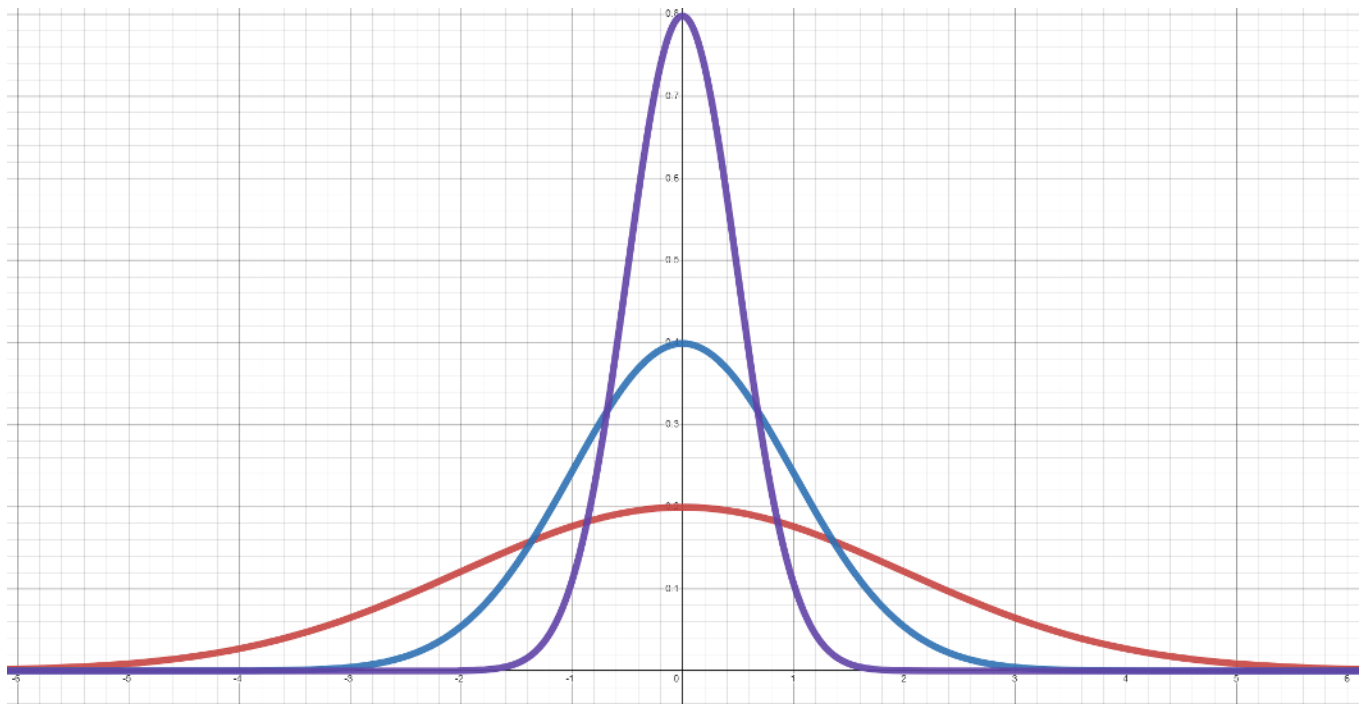
Graphing the density of the $\mathcal{N}(0, 4)$ random variable yields:



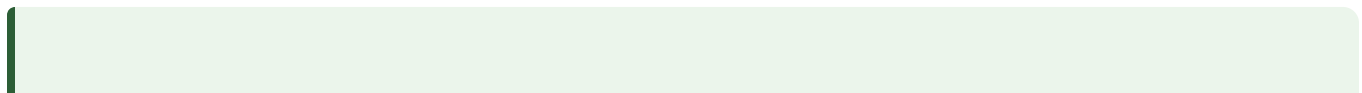
Intuitively, the second parameter represents the thickness and since the second parameter increased, the distribution is now thicker. Finally, graphing the density of the $\mathcal{N}(0, 0.25)$ random variable random variables yields:



Intuitively, the second parameter represents the thickness and since the second parameter decreased, the distribution is now thinner. To get a side by side comparison, let us draw the densities of the $\mathcal{N}(0, 1)$, $\mathcal{N}(0, 4)$ and $\mathcal{N}(0, 0.25)$ all on one graph:



Now that we have a feel for how the two parameters affect the graph of the density, we present the following theorem.



Theorem 11.4.1

Theorem: If $X \sim \mathcal{N}(\mu, \sigma^2)$ then

$$\mathbb{E}[X] = \mu \text{ and } \text{Var}[X] = \sigma^2$$

We have considered five different normal random variables in the above discussion. However, there are infinitely many normal random variables we can have. Among the infinitely many of them, there is one normal random variable that we designate to be *the standard*, and so we will single out this particular normal random variable and we will attach a special name to it. In fact, to stress its distinction, we will assign this specific normal random variable its own letter and we will give the pdf and cdf special letters too. Recall that most times we denote a random variable by X , the density by f and the cdf by F . But since this random variable is so important, we will denote it by Z and we denote the density by the Greek letter ϕ (lowercase phi) and the cdf by the Greek letter Φ (uppercase phi).

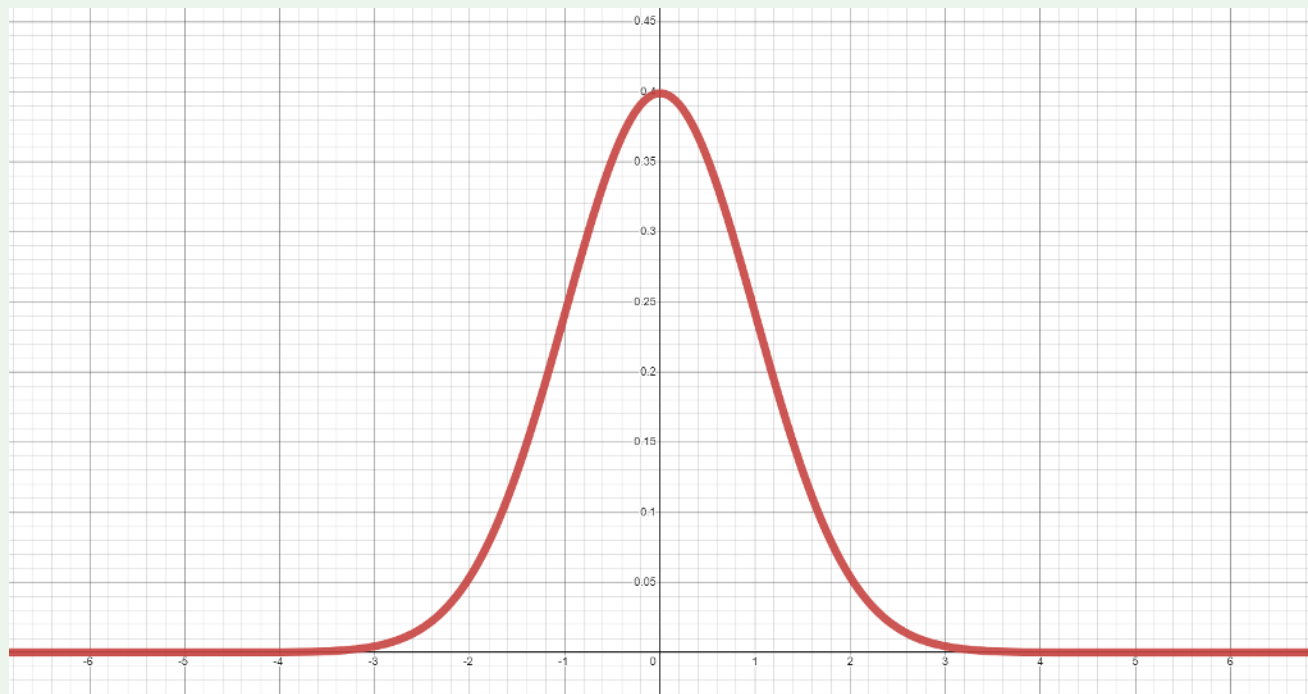
Definition: The Standard Normal Random Variable

Definition: We call the $\mathcal{N}(0, 1)$ the **standard normal random variable** and write $Z \sim \mathcal{N}(0, 1)$. Its density is given by

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \text{ if } -\infty < z < \infty \quad (11.4.3)$$

and its cdf is given by

$$\Phi(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt \quad (11.4.4)$$



With this in mind, allow us to consider the following example that I would like to **attempt** to answer in three different ways.

✓ Example (Method 1) 11.4.1

Suppose $Z \sim \mathcal{N}(0, 1)$.

Find $P(Z \leq 1)$.

Find $P(Z \geq 1.5)$.

Find $P(-1.36 \leq Z \leq 1.72)$.

First Attempt at a Solution:

Allow us to try to use the definition of a continuous random variable to answer this problem. By the definition of a continuous random variable, we see that

$$P(Z \leq 1) = \Phi(1) = \int_{-\infty}^1 \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt = \lim_{t \rightarrow -\infty} \left[\int_t^1 \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt \right]$$

Can we find an antiderivative of $\frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$? Unfortunately, the answer is no. Although the antiderivative exists, this antiderivative is not an elementary function which means no matter how hard we try, or what integration technique we use, we will never be able to produce the antiderivative of $\frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$ in terms of functions we know. (We can, however, express the $\int \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$ as an infinite series).

$$P(Z \leq 1) = \Phi(1) = \int_{-\infty}^1 \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt = \lim_{t \rightarrow -\infty} \left[\int_t^1 \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt \right] = ???$$

The point remains that we cannot perform the above computation by hand. A similar story holds for the following two parts of the above example.

$$P(Z \geq 1.5) = 1 - P(Z < 1.5) = 1 - P(Z \leq 1.5) = 1 - \Phi(1.5) = 1 - \int_{-\infty}^{1.5} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt = ???$$

$$P(-1.36 \leq Z \leq 1.72) = \Phi(1.72) - \Phi(-1.36) = \int_{-1.36}^{1.72} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt = ???$$

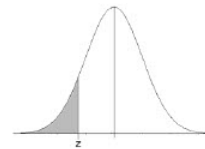
In summary, we are stuck since in each case, we run into an integral which we cannot evaluate and so we fail to obtain a numerical answer.

We have failed our first attempt at the above problem and so the question now becomes, is there a way we can get around this?

Luckily, the answer is yes! Specifically, what we can do is the following: For ONLY the standard normal random variable, we can create a chart which will give us the cdf, $\Phi(z)$, for important/common values. Of course, this chart has to be compromise. What do I mean by this? You see, the support of Z has infinitely many real numbers. Since there are infinitely many real numbers, then technically we would need an infinite chart which is not exactly practical. Thus, our chart will give the CDF for only specific/important/common values of the standard normal random variable. Using this chart is Method 2:

Standard Normal Cumulative Probability Table

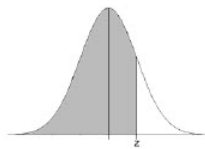
Cumulative probabilities for NEGATIVE z-values are shown in the following table:



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

Standard Normal Cumulative Probability Table

Cumulative probabilities for POSITIVE z-values are shown in the following table:



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177

1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Chart provided by: <https://www.math.arizona.edu/~jwatkins/normal-table.pdf>

✓ Example (Method 2) 11.4.1

Suppose $Z \sim \mathcal{N}(0, 1)$.

Find $P(Z \leq 1)$.

Find $P(Z \geq 1.5)$.

Find $P(-1.36 \leq Z \leq 1.72)$.

Second Attempt at a Solution:

We now use the standard normal cdf chart presented above.

$$P(Z \leq 1) = \Phi(1) = \Phi(1.00) \approx 0.8413$$

$$P(Z \geq 1.5) = 1 - P(Z < 1.5) = 1 - P(Z \leq 1.5) = 1 - \Phi(1.5) = 1 - \Phi(1.50) \approx 1 - 0.9332 = 0.0668$$

$$P(-1.36 \leq Z \leq 1.72) = \Phi(1.72) - \Phi(-1.36) \approx 0.9573 - 0.0869 = 0.8704$$

And so we were successful in the above attempt. This approach is what we did in the 1700s/1800s. However, our third method uses a more modern approach - making use of the the calculator commands!

✓ Example (Attempt 3) 11.4.1

Suppose $Z \sim \mathcal{N}(0, 1)$.

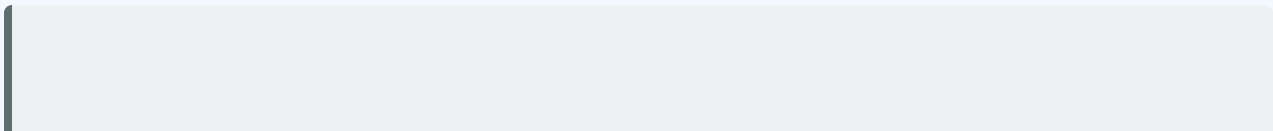
Find $P(Z \leq 1)$.

Find $P(Z \geq 1.5)$.

Find $P(-1.36 \leq Z \leq 1.72)$.

Third Attempt at a Solution:

The TI-84 is equipped with a normalcdf command. To use it, we note the following:



Note

If $X \sim \mathcal{N}(\mu, \sigma^2)$ then:

$$1) P(a \leq X \leq b) = \int_a^b \phi(z) dz = \text{normalcdf}(a, b, \mu, \sigma)$$

$$2) P(X \leq a) = \int_{-\infty}^a \phi(z) dz = \text{normalcdf}(-1E99, a, \mu, \sigma)$$

where $-1E99$ is how your calculator approximates $-\infty$.

Hence,

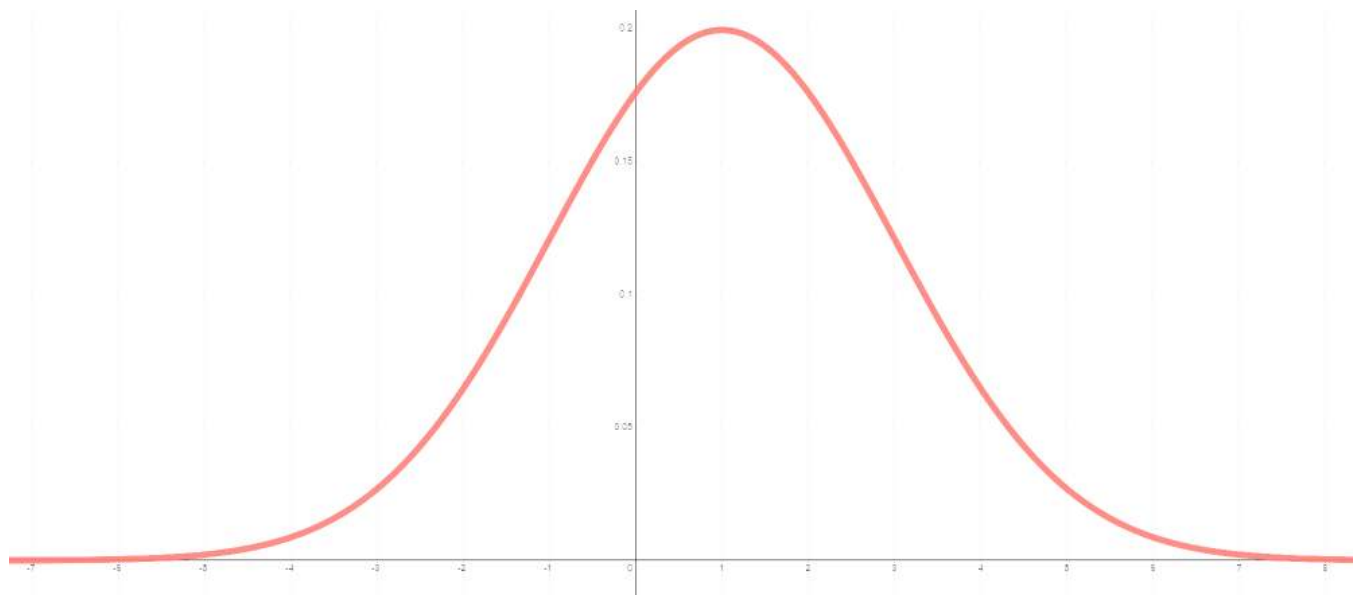
$$P(Z \leq 1) = \text{normalcdf}(-1E99, 1, 0, 1) \approx 0.8413447404$$

$$P(Z \geq 1.5) = 1 - P(Z < 1.5) = 1 - P(Z \leq 1.5) = 1 - \text{normalcdf}(-1E99, 1.5, 0, 1) \approx 0.668072287$$

$$P(-1.36 \leq Z \leq 1.72) = \text{normalcdf}(-1E99, a, \mu, \sigma) \approx 0.8703678946$$

Thus, given the standard normal random variable, I can find any probabilities associated with it. But what if we did not necessarily have the standard normal random variable? For instance, what if I said $X \sim \mathcal{N}(1, 4)$. Find $F(3) = P(X \leq 3)$. How could we do this?

We *could* do attempt like we did in Attempt 1. That is, in theory, we know the answer should be the integral of the density function over the appropriate region. Geometrically, we see the density looks like:



And so

$$P(X \leq 3) = \int_{-\infty}^3 \frac{1}{2\sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{t-1}{2} \right)^2} dt = \text{???}$$

And again, this becomes a function which we cannot integrate by hand.

Alternatively, we *could* develop a chart for the cumulative distribution function of the $\mathcal{N}(1, 4)$ but does this sound practical? If every time we had a normal random variable, we had to lookup a chart, then how many different charts would we have? Infinitely many of them!

And so we need a more efficient method. The quick answer is to use the modern approach with our calculators:

✓ Example 11.4.1

If $X \sim \mathcal{N}(1, 4)$, then find $P(X \leq 3)$.

Answer

$$P(X \leq 3) = \text{normalcdf}(-1E99, 3, 1, 2) \approx 0.8413447404.$$

However, how did mathematicians and statisticians in the 1800s answer this question when they did not have a calculator readily available to them? The answer comes in the form of the following theorem:

🔗 Theorem: The Standardization Process 11.4.1

Theorem: If $X \sim \mathcal{N}(\mu, \sigma^2)$ then

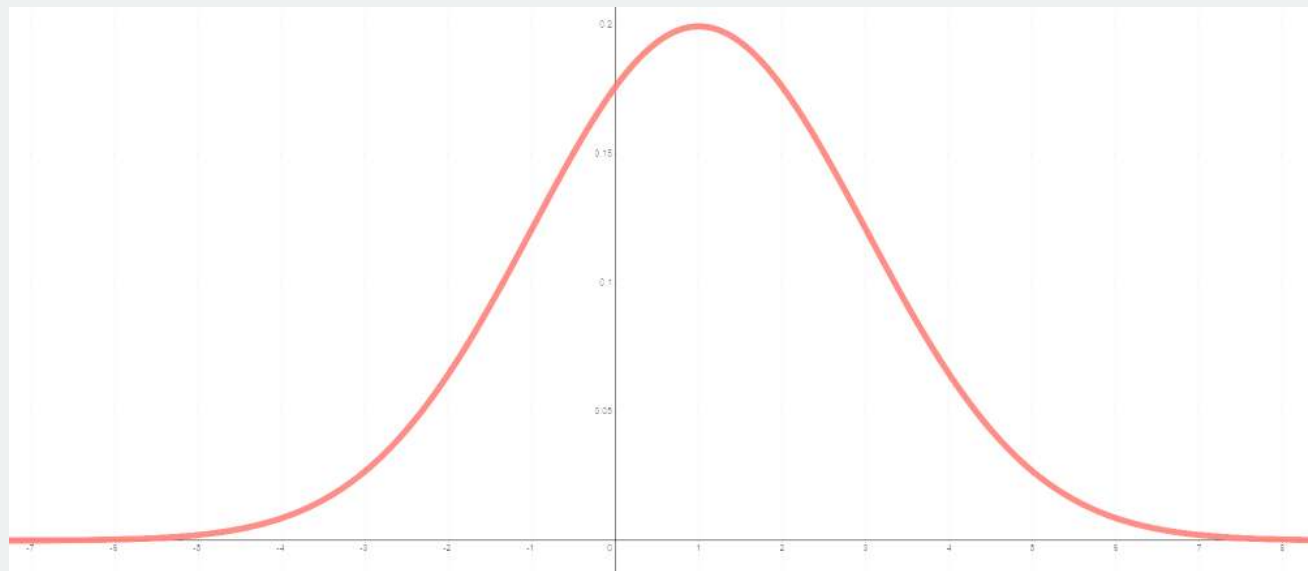
$$\frac{X - \mu}{\sigma} = Z$$

Furthermore, if F denotes the cdf of X , then

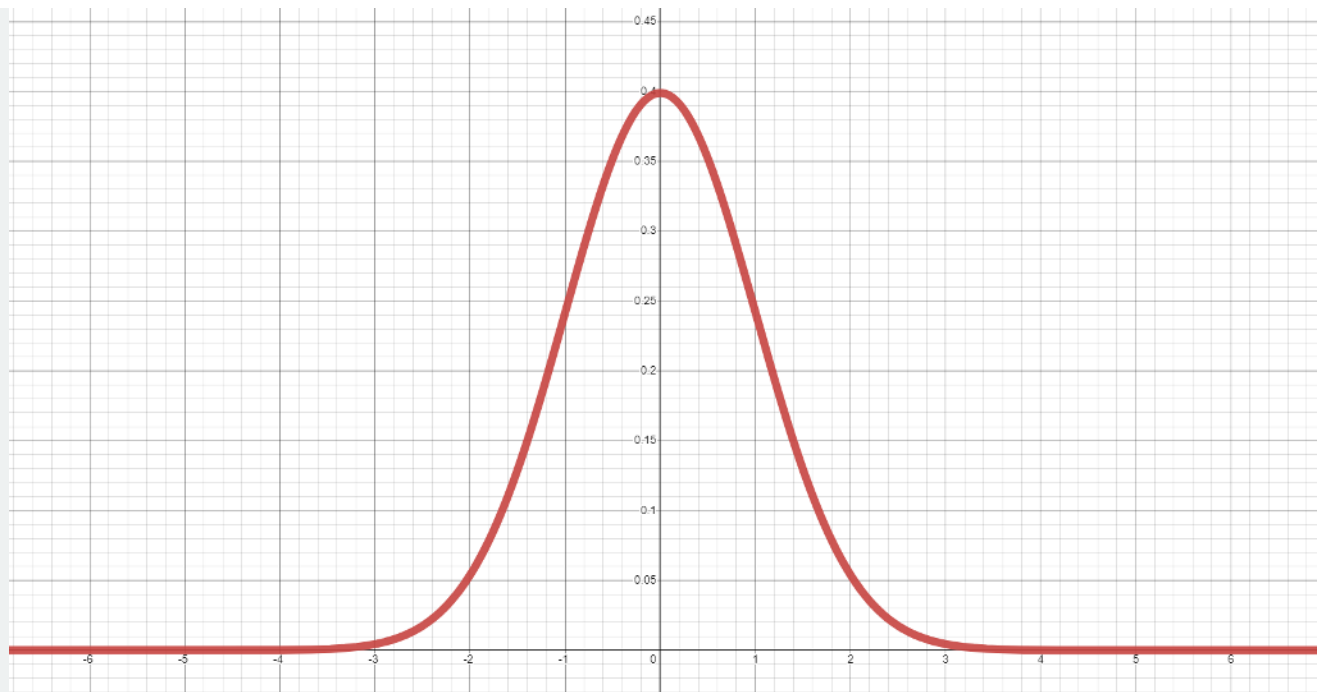
$$F(x) = \Phi\left(\frac{X - \mu}{\sigma}\right)$$

📌 Interpretation of the Standardization Process

Let us suppose we start with the $\mathcal{N}(1, 4)$ random variable as pictured below:



If we were to apply the following linear transformation to X , $\frac{X-1}{2}$, then the random variable becomes



which is precisely Z , the standard normal random variable!

Moreover, this is true for *any* normal random variable. If we start with any normal random variable X with parameters μ and σ^2 , then the random variable $\frac{X-\mu}{\sigma}$ becomes Z .

In summary, given any normal random variable, if we subtract the first parameter and divide by the square root of the second parameter, then we will obtain the standard normal random variable, Z .

Using the above theorem, allow us to answer the question.

✓ Example 11.4.1

If $X \sim \mathcal{N}(1, 4)$, then find $P(X \leq 3)$.

Answer

$$\begin{aligned} P(X \leq 3) &= P\left(\frac{X-1}{2} \leq \frac{3-1}{2}\right) \\ &= P(Z \leq 1) \\ &= \Phi(1) \text{ via the standard normal table} \\ &\approx 0.8413 \end{aligned}$$

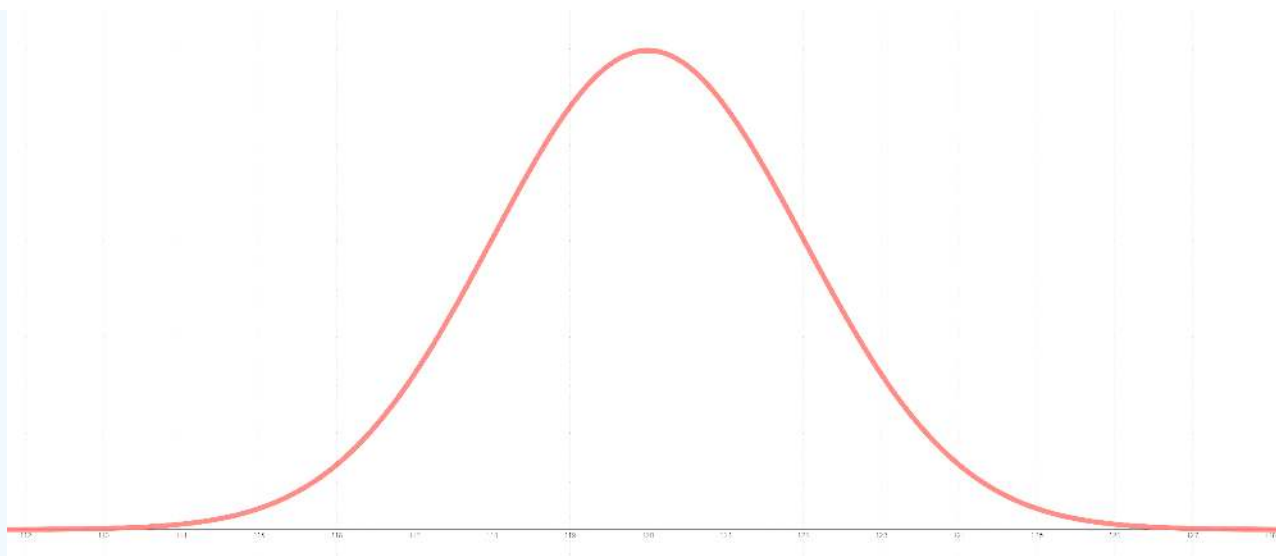
To summarize, these techniques, allow us to take a look at one final example.

✓ Example 11.4.1

If $X \sim \mathcal{N}(120, 2^2)$, then find $P(116 \leq X \leq 118)$.

Answer

Before we answer this, let us understand the picture: The $\mathcal{N}(120, 2^2)$ looks like the following:



Here are the four ways we can attempt to answer this problem, of which the last three attempts are preferred.

1) Of course we could say

$$P(116 \leq X \leq 118) = \int_{116}^{118} \frac{1}{2\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{t-120}{2}\right)^2} dt$$

but then we would not be able to integrate this.

2) Alternatively we can perform the simple calculator command:

$$P(116 \leq X \leq 118) = F(118) - F(116) = \text{normalcdf}(116, 118, 120, 2) \approx 0.1359051975$$

3) Or we may use the standardization process with the calculator.

$$\begin{aligned} P(116 \leq X \leq 118) &= P\left(\frac{116-120}{2} \leq \frac{X-120}{2} \leq \frac{118-120}{2}\right) \\ &= P(-2 \leq Z \leq -1) \\ &= \Phi(-1) - \Phi(-2) \\ &= \text{normalcdf}(-2, -1, 0, 1) \\ &\approx 0.1359051975 \end{aligned}$$

4) Or we may use the standardization process with the calculator.

$$\begin{aligned} P(116 \leq X \leq 118) &= P\left(\frac{116-120}{2} \leq \frac{X-120}{2} \leq \frac{118-120}{2}\right) \\ &= P(-2 \leq Z \leq -1) \\ &= \Phi(-1) - \Phi(-2) \\ &\approx 0.1587 - 0.0228 \\ &= 0.1359 \end{aligned}$$

11.4: The Normal Distribution is shared under a [CC BY](https://creativecommons.org/licenses/by/4.0/) license and was authored, remixed, and/or curated by LibreTexts.

CHAPTER OVERVIEW

Chapter 12: Week 12

[12.1: The Central Limit Theorem](#)

[Chapter 12: Week 12](#) is shared under a [CC BY](#) license and was authored, remixed, and/or curated by LibreTexts.

12.1: The Central Limit Theorem

In the last section, we introduced the normal random variable and primarily discussed how to find the probabilities associated with a normal random variable. We also mentioned that the normal random variable is arguably the most important continuous random variable. We shall now find out why. We first make the following definition:

Definition: Sample Mean

Definition: Let $X_1, X_2, X_3, \dots, X_n$ be random variables. We call the new random variable,

$$\frac{X_1 + X_2 + X_3 + \dots + X_n}{n}$$

the **sample mean** and we denote it by $\bar{X}_n$.

✓ Example 12.1.1

- 1) Suppose X_i models the outcome of our i^{th} roll of a fair die. Then the random variable $\frac{X_1 + X_2 + \dots + X_{100}}{100}$ is the sample mean of the 100 rolls.
- 2) Let us suppose there are 30 people in a classroom and let X_i models the amount of time person i spends watching TV on a daily basis. Then the random variable $\frac{X_1 + X_2 + \dots + X_{30}}{30}$ represents the sample mean.

The question we will consider is if the random variables, $X_1, X_2, X_3, \dots, X_n$ are all independent and identically distributed, then can we say anything about their sample mean? The answer is yes and the details are given to us in one of the most important theorems in all of probability and statistics - The Central Limit Theorem.

Theorem: The Central Limit Theorem 12.1.2

Theorem: Suppose the random variables $X_1, X_2, X_3, \dots, X_n$ are all independent and identically distributed where each random variable has mean μ and variance σ^2 . Then for each fixed x ,

$$\lim_{n \rightarrow \infty} P \left[\frac{\bar{X}_n - \mu}{\sigma / \sqrt{n}} \leq x \right] = \Phi(x)$$

Or equivalently

$$\lim_{n \rightarrow \infty} P \left[\frac{X_1 + X_2 + \dots + X_n - n\mu}{\sigma \sqrt{n}} \leq x \right] = \Phi(x)$$

Interpretation of The Central Limit Theorem:

Suppose we have a bunch of random variables, $X_1, X_2, X_3, \dots, X_n$ that are all independent and have the same distribution. (For instance, all of them are Uniform random variables with the same parameters or all of them are Exponential random variables with the same parameters). Since they come from the same or identical distribution, then they all have the same average, μ , and same variance, σ^2 . The Central Limit Theorem tells us that:

- 1) the new random variable, $\frac{X_1 + X_2 + \dots + X_n}{n} = \bar{X}_n$ will approximately be $\mathcal{N}(\mu, \frac{\sigma^2}{n})$.
- 2) the new random variable, $X_1 + X_2 + \dots + X_n$ will be approximately $\mathcal{N}(n\mu, n\sigma^2)$.

Additionally, notice how general the Central Limit Theorem is! We are saying the distribution of $X_1, X_2, X_3, \dots, X_n$ can be anything we wish, as long as the random variables all have the same distribution. Regardless, the sample mean (as well as the sum) will tend to a normal distribution.

Before we consider our examples, allow us to take a few moments to use the following applet. First select a distribution and then select the value of n . Notice that regardless of the starting distribution, the sample mean will tend to a normal distribution as n increases.

📌 Discover the central limit theorem activity

Uniform

Skewed Left

Skewed Right

Normal

User

Select the distribution that you want to sample from.

✓ Example 12.1.3

Let $\bar{X}_{25}$ denote the sample mean of 25 random variables where each random variable comes from a distribution with a mean of 15 and variance of 4. Find $P(14.4 \leq \bar{X}_{25} \leq 15.6)$.

Answer

In this problem $\mu = 15$ and $\sigma^2 = 4$. By the Central Limit Theorem, $\bar{X}_{25}$ will approximately be

$$\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) = \mathcal{N}\left(15, \frac{4}{25}\right)$$

To finish the problem, we note that

$$P(14.4 \leq \bar{X}_{25} \leq 15.6) = \text{normalcdf}(-1E99, 15.6, 15, 2/5) - \text{normalcdf}(-1E99, 14.4, 15, 2/5) = 0.8663855426$$

If you do not have a calculator, then you are forced to use the theorem from last section with the CDF chart:

$$\begin{aligned}
 P(14.4 \leq \bar{X}_{25} \leq 15.6) &= P\left(\frac{14.4 - 15}{2/5} \leq \frac{X - 15}{2/5} \leq \frac{15.6 - 15}{2/5}\right) \\
 &\approx P(-1.5 \leq Z \leq 1.5) \\
 &= 0.9332 - 0.0668 \\
 &= 0.8664
 \end{aligned}$$

✓ Example 12.1.4

Let $\bar{X}_{25}$ denote the sample mean when the underlying random variables come from a distribution whose pdf is given by

$$f(x) = \begin{cases} \frac{x^3}{4} & \text{if } 0 < x < 2 \\ 0 & \text{otherwise} \end{cases}$$

Find/approximate $P(1.5 \leq \bar{X}_{25} \leq 1.65)$.

Answer

Before we can even answer this question, we have to understand what the problem is asking us. This question is saying suppose you have these random variables, $X_1, X_2, \dots, X_{25}$ where these random variables all have the above density. We are asked to find or approximate

$$P\left(1.5 \leq \frac{X_1 + X_2 + \dots + X_{25}}{25} \leq 1.65\right)$$

By the Central Limit Theorem, $\frac{X_1 + X_2 + \dots + X_{25}}{25} = \bar{X}_{25}$ will approximately be $\mathcal{N}(\mu, \frac{\sigma^2}{25})$. Let us find μ and σ^2 .

$$\mu = \mathbb{E}[X_i] = \int_{\text{all } x} x f(x) dx = \int_0^2 x \left(\frac{x^3}{4}\right) dx = \frac{8}{5} = 1.6$$

$$\begin{aligned}
 \sigma^2 &= \text{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \\
 &= \mathbb{E}[X^2] - 1.6^2 \\
 &= \int_0^2 x^2 \left(\frac{x^3}{4}\right) dx - 1.6^2 \\
 &= \frac{8}{3} - 1.6^2 \\
 &= \frac{8}{75}
 \end{aligned}$$

By the Central Limit Theorem, $\frac{X_1 + X_2 + \dots + X_{25}}{25} = \bar{X}_{25}$ will approximately be $\mathcal{N}(\mu, \frac{\sigma^2}{25})$ which turns out to be $\mathcal{N}(1.6, \frac{8/75}{25}) = \mathcal{N}(1.6, \frac{8}{1875})$.

So $\bar{X}_{25} \sim \mathcal{N}(1.6, \frac{8}{1875})$. Hence,

$$P(1.5 \leq \bar{X}_{25} \leq 1.65) \approx \text{normalcdf}(1.5, 1.65, 1.6, \sqrt{\frac{8}{1875}}) \approx 0.7151096523$$

Or alternatively, if we wish to use the standardization process, we obtain the following:

$$\begin{aligned}
 P(1.5 \leq \bar{X}_{25} \leq 1.65) &= P\left(\frac{1.5 - 1.6}{\sqrt{\frac{8}{1875}}} \leq \frac{\bar{X}_{25} - 1.6}{\sqrt{\frac{8}{1875}}} \leq \frac{1.65 - 1.6}{\sqrt{\frac{8}{1875}}}\right) \\
 &\approx P(-1.531 \leq Z \leq 0.765) \\
 &\approx P(-1.53 \leq Z \leq 0.77) \\
 &= \Phi(0.77) - \Phi(-1.53) \\
 &= 0.7794 - 0.0630 \\
 &= 0.7164
 \end{aligned}$$

✓ Example 12.1.5

Suppose $X_1, X_2, \dots, X_{20}$ are all independent and identically distributed as $\mathcal{U}[0, 1]$. Find $P(X_1 + X_2 + \dots + X_{20} < 9.1)$.

Answer

Recall if $X \sim \mathcal{U}[a, b]$ then $\mathbb{E}[X] = \frac{a+b}{2}$ and $\mathbb{V}ar[X] = \frac{(b-a)^2}{12}$. Hence, for the $\mathcal{U}[0, 1]$, $\mathbb{E}[X] = \frac{0+1}{2} = \frac{1}{2} = \mu$ and $\mathbb{V}ar[X] = \frac{(1-0)^2}{12} = \frac{1}{12} = \sigma^2$.

We know by The Central Limit Theorem, $X_1 + X_2 + \dots + X_n$ will be approximately $\mathcal{N}(n\mu, n\sigma^2)$ and so $X_1 + X_2 + \dots + X_{20}$ will be approximately $\mathcal{N}(20(\frac{1}{2}), 20(\frac{1}{12})) = \mathcal{N}(10, \frac{5}{3})$. Thus, we have the following:

$$\begin{aligned}
 P(X_1 + X_2 + \dots + X_{20} < 9.1) &= normalcdf\left(-1E99, 9.1, 10, \sqrt{\frac{5}{3}}\right) \\
 &\approx 0.2428584558
 \end{aligned}$$

This page titled [12.1: The Central Limit Theorem](#) is shared under a [CC BY](#) license and was authored, remixed, and/or curated by [Larry Green](#).

CHAPTER OVERVIEW

Chapter 13: Week 13

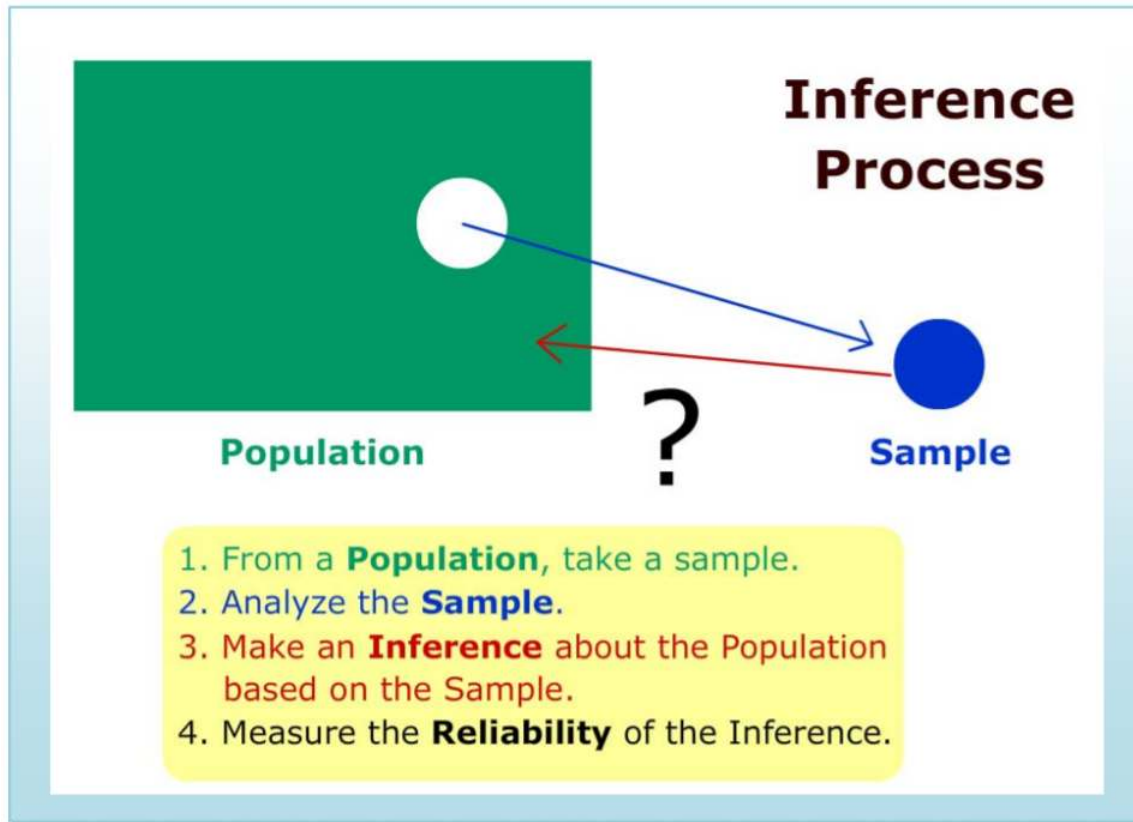
[13.1: Inferential Statistics](#)

[13.2: Confidence Intervals](#)

Chapter 13: Week 13 is shared under a [CC BY](#) license and was authored, remixed, and/or curated by LibreTexts.

13.1: Inferential Statistics

The reason we conduct statistical research is to obtain an understanding about phenomena in a population. For example, we may want to know if a potential drug is effective in treating a disease. Since it is not feasible or ethical to distribute an experimental drug to the entire population, we instead must study a small subset of the population called a sample. We then analyze the sample and make an inference about the population based on the sample. Using probability theory and the Central Limit Theorem, we can then measure the reliability of the inference.



Example: Home value appraisal

Lupe is trying to sell her house and needs to determine the market value of the home. The population in this example would be all the homes that are similar to hers in the neighborhood.

Sampled Homes Adjusted Sales Price		
\$420,000	\$440,000	\$470,000
\$430,000	\$450,000	\$470,000
\$430,000	\$460,000	\$480,000

Lupe's realtor chooses for the **sample** nine recent homes in this neighborhood that sold in the last six months. The realtor then adjusts some of the sales prices to account for differences among Lupe's home and the sold homes.

Next the realtor takes the mean of the adjusted sample and recommends to Lupe a market value for Lupe's home of \$450,000. The realtor has made an **inference** about the mean value of the population.

To measure the **reliability** of the inference, the realtor should look at factors such as: the small sample size, changes in values of homes over the last six months, or the fact that Lupe's home is not exactly like the sampled homes.

This page titled [13.1: Inferential Statistics](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Maurice A. Geraghty](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

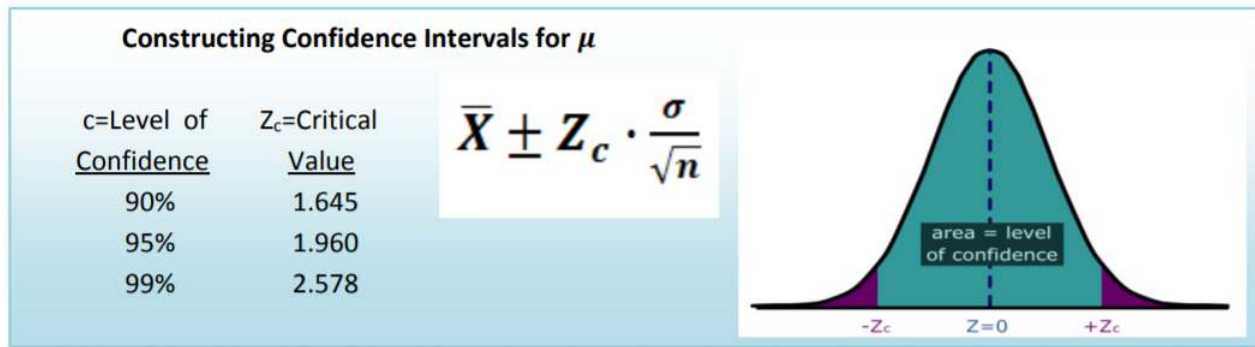
- **9.1: Inferential Statistics** by [Maurice A. Geraghty](#) is licensed [CC BY-SA 4.0](#). Original source: <http://nebula2.deanza.edu/~mo/holisticInference.html>.

13.2: Confidence Intervals

Using probability and the Central Limit Theorem, we can design an Interval Estimate called a **Confidence Interval** which has a known probability (**Level of Confidence**) of capturing the true population parameter.

Confidence Interval for Population Mean

To find a confidence interval for the population mean (μ) when the population standard deviation (σ) is known, and n is sufficiently large, we can use the Standard Normal Distribution probability distribution function to calculate the critical values for the Level of Confidence:



Example: Students working

The Dean wants to estimate the mean number of hours that students worked per week. A sample of 49 students showed a mean of 24 hours with a standard deviation of 4 hours. The point estimate is 24 hours (sample mean). What is the 95% confidence interval for the average number of hours worked per week by the students?

Solution

$$24 \pm \frac{1.96 \cdot 4}{\sqrt{49}} = 24 \pm 1.12 = (22.88, 25.12) \text{ hours per week}$$

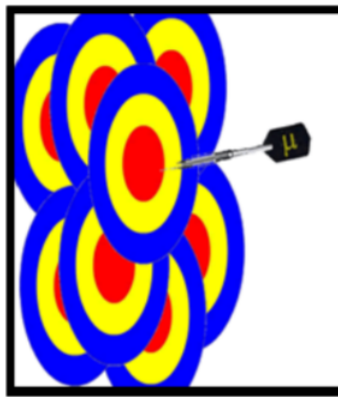
The margin of error for the confidence interval is 1.12 hours. We can say with 95% confidence that the mean number of hours worked by students is between 22.88 and 25.12 hours per week.

If the level of confidence is increased, then the margin of error will also increase. For example, if we increase the level of confidence to 99% for the above example, then:

$$24 \pm \frac{2.578 \cdot 4}{\sqrt{49}} = 24 \pm 1.47 = (22.53, 25.47) \text{ hours per week}$$

Some important points about Confidence Intervals

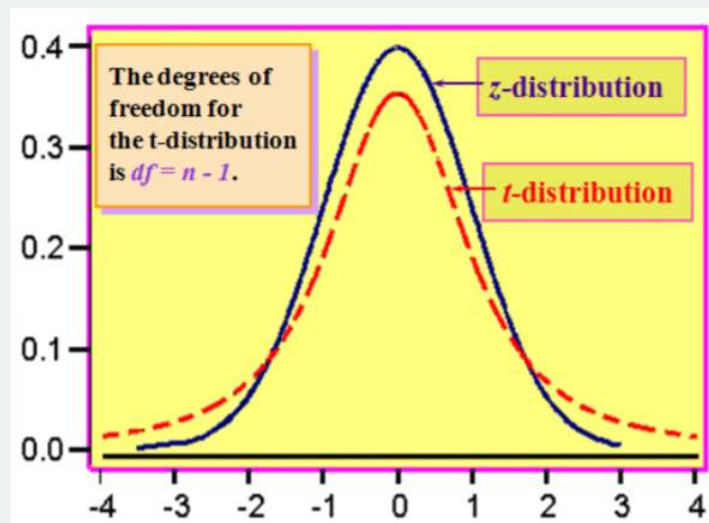
- The confidence interval is constructed from random variables calculated from sample data and attempts to predict an unknown but fixed population parameter with a certain level of confidence.
- Increasing the level of confidence will always increase the margin of error.
- It is impossible to construct a 100% Confidence Interval without taking a census of the entire population.
- Think of the population mean as a dart that always goes to the same spot, and the confidence interval as a moving target that tries to "catch the dart." A 95% confidence interval would be like a target that has a 95% chance of catching the dart.



Confidence Interval for Population Mean using Sample Standard Deviation – Student's t Distribution

The formula for the confidence interval for the mean requires the knowledge of the population standard deviation (σ). In most real-life problems, we do not know this value for the same reasons that we do not know the population mean. This problem was solved by the Irish statistician William Sealy Gosset, an employee at Guinness Brewing. Gosset, however, was prohibited by Guinness from using his own name in publishing scientific papers. He published under the name "A Student", and therefore the distribution he discovered was named "Student's t -distribution"⁷¹.

Characteristics of Student's t Distribution



- It is continuous, bell-shaped, and symmetrical about zero like the z distribution.
- There is a **family** of t -distributions sharing a mean of zero but having different standard deviations based on **degrees of freedom**.
- The t -distribution is more spread out and flatter at the center than the Z -distribution, but approaches the Z -distribution as the sample size gets larger.

Confidence Interval for μ

$$\bar{X} \pm t_c \frac{s}{\sqrt{n}} \text{ with degrees of freedom} = n - 1$$

Example: Rating health care plans

Last year Sally belonged to an Health Maintenance Organization (HMO) health care plan that had a population average rating of 62 (on a scale from 0-100, with '100' being best); this was based on records accumulated about the HMO over a long period of time. This year Sally switched to a new HMO. To assess the population mean rating of the new HMO, 20 members of this

HMO are polled and they give the HMO an average rating of 65 with a standard deviation of 10. Find and interpret a 95% confidence interval for population average rating of the new HMO.

Solution

The t distribution will have $20-1 = 19$ degrees of freedom. Using a table or technology, the critical value for the 95% confidence interval will be $t_c = 2.093$

$$65 \pm \frac{2.093 \cdot 10}{\sqrt{20}} = 65 \pm 4.68 = (60.32, 69.68) \text{ HMO rating}$$

With 95% confidence we can say that the rating of Sally's new HMO is between 60.32 and 69.68. Since the quantity 62 is in the confidence interval, we cannot say with 95% certainty that the new HMO is either better or worse than the previous HMO.

Confidence Interval for Population Proportion

Recall from the section on random variables the binomial distribution where p represented the proportion of successes in the population. The binomial model was analogous to coin-flipping, or yes/no question polling. In practice, we want to use sample statistics to estimate the population proportion (p).

The sample proportion ($\hat{p}$) is the proportion of successes in the sample of size n and is the point estimator for p . Under the Central Limit Theorem, if $np > 10$ and $n(1-p) > 10$, the distribution of the sample proportion $\hat{p}$ will have an approximately Normal Distribution.

Normal Distribution for $\hat{p}$ if Central Limit Theorem conditions are met.

$$\mu_{\hat{p}} = p \quad \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Using this information we can construct a confidence interval for p , the population proportion:

Confidence interval for p

$$\hat{p} \pm Z \sqrt{\frac{p(1-p)}{n}} \approx \hat{p} \pm Z \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Example: Talking and driving

200 California drivers were randomly sampled and it was discovered that 25 of these drivers were illegally talking on their cell phones without the use of a hands-free device. Find the point estimator for the proportion of drivers who are using their cell phones illegally and construct a 99% confidence interval.



Solution

The point estimator for p is $\hat{p} = \frac{25}{200} = .125$ or 12.5%.

A 99% confidence interval for p is: $0.125 \pm 2.576 \sqrt{\frac{.125(1 - .125)}{200}} = .125 \pm .060$

The margin of error for this poll is 6% and we can say with 99% confidence that the true percentage of drivers who are using their cell phones illegally is between 6.5% and 18.5%

Point Estimator for Population Standard Deviation

We often want to study the variability, volatility or consistency of a population. For example, two investments both have expected earnings of 6% per year, but one investment is much riskier, with higher ups and downs. To estimate variation or volatility of a data set, we will use the sample standard deviation s as a point estimator of the population standard deviation (σ).

Example

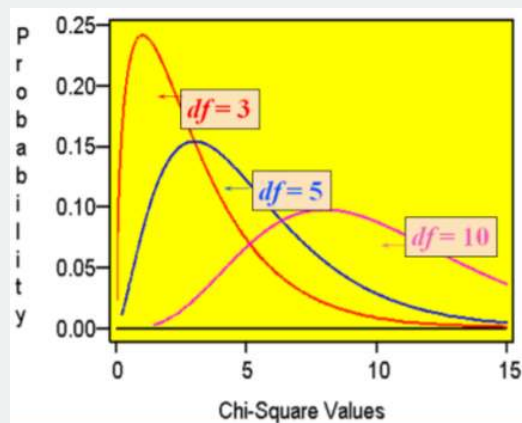
Investments A and B are both known to have a rate of return of 6% per year. Over the last 24 months, Investment A has a sample standard deviation of 3% per month, while Investment B has a sample standard deviation of 5% per month. We would say that Investment B is more volatile and riskier than Investment A due to the higher estimate of the standard deviation.

To create a confidence interval for an estimate of standard deviation, we need to introduce a new distribution, called the Chi-square (χ^2) distribution.

The Chi-square χ^2 Distribution

The Chi-square distribution is a family of distributions related to the Normal Distribution, since it represents a sum of independent squared standard Normal Random Variables. Like the Student's t distribution, the degrees of freedom will be $n - 1$ and will determine the shape of the distribution. Also, since the Chi-square represents squared data, the inference will be about the variance rather than about the standard deviation.

Characteristics of Chi-square χ^2 Distribution



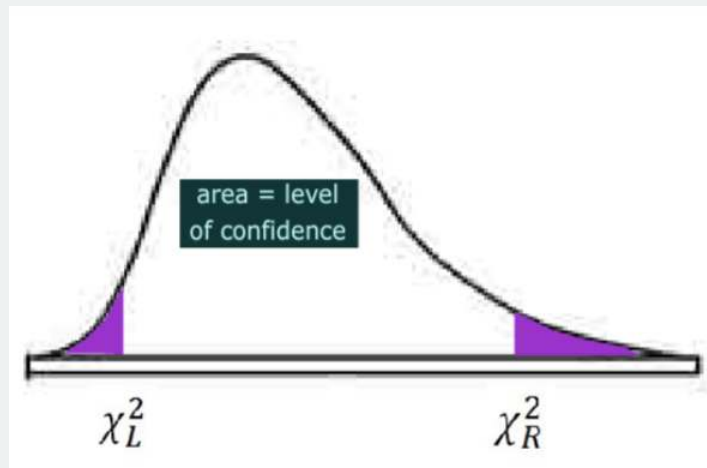
- It is positively skewed
- It is non-negative
- It is based on degrees of freedom ($n - 1$)
- When the degrees of freedom change, a new distribution is created $\frac{(n - 1)s^2}{\sigma^2}$ will have Chi-square distribution.

Confidence Interval for Population Variance and Standard Deviation

Since the Chi-square represents **squared data**, we can construct confidence intervals for the population variance (σ^2), and take the square root of the endpoints to get a confidence interval for the population standard deviation. Due to the skewness of the Chi-

square distribution the resulting confidence interval will not be centered at the point estimator, so the margin of error form used in the prior confidence intervals doesn't make sense here.

Confidence Interval for population variance (σ^2)



- Confidence is **NOT** symmetric since chi-square distribution is not symmetric.
- Take square root of both endpoints to get confidence interval for the population standard deviation (σ).

$$\left(\frac{(n-1)s^2}{\chi_R^2}, \frac{(n-1)s^2}{\chi_L^2} \right)$$

Example: Performance risk in finance

In performance measurement of investments, standard deviation is a measure of volatility or risk. Twenty monthly returns from a mutual fund show an average monthly return of 1 percent and a sample standard deviation of 5 percent. Find a 95% confidence interval for the monthly standard deviation of the mutual fund.



Solution

The Chi-square distribution will have $20-1 = 19$ degrees of freedom. Using technology, we find that the two critical values are $\chi_L^2 = 8.90655$ and $\chi_R^2 = 32.8523$

Formula for confidence interval for σ is: $\left(\sqrt{\frac{(19)5^2}{32.8523}}, \sqrt{\frac{(19)5^2}{8.90655}} \right) = (3.8, 7.3)$

One can say with 95% confidence that the standard deviation for this mutual fund is between 3.8 and 7.3 percent per month.

This page titled [13.2: Confidence Intervals](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Maurice A. Geraghty](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- **9.3: Confidence Intervals** by [Maurice A. Geraghty](#) is licensed [CC BY-SA 4.0](#). Original source: <http://nebula2.deanza.edu/~mo/holisticInference.html>.

CHAPTER OVERVIEW

Chapter 14: Week 14

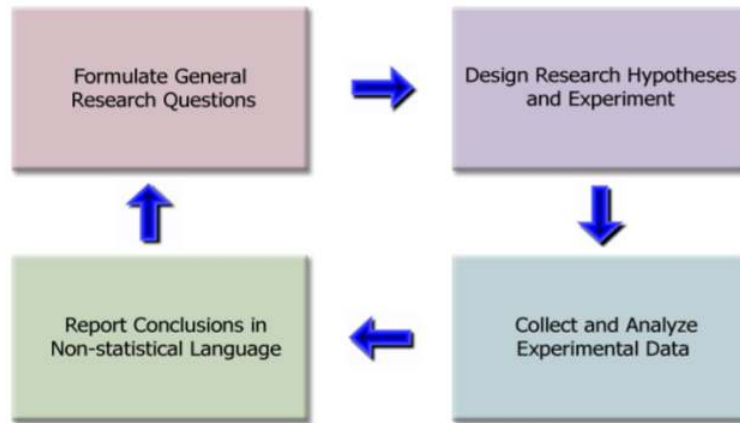
In the prior section we used statistical inference to make an estimate of a population parameter and to measure the reliability of the estimate through a confidence interval. In this section, we will explore in detail the use of statistical inference in testing a claim about a population parameter; this inference is the heart of the scientific method used in research.

- 14.1: Procedures of Hypotheses Testing and the Scientific Method
- 14.2: Formulate General Research Questions
- 14.3: Design Research Hypotheses and Experiment
- 14.4: Collect and Analyze Experimental Data
- 14.5: Report Conclusions in Non-statistical Language
- 14.6: Test of Mean vs. Hypothesized Value – A Complete Example
- 14.7: Type II Error and Statistical Power
- 14.8: New Models for One Population Inference, Similar Procedures
- 14.9: The p-value- Misconceptions and Proper Usage

This page titled [Chapter 14: Week 14](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Maurice A. Geraghty](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

14.1: Procedures of Hypotheses Testing and the Scientific Method

The actual conducting of a hypothesis test is only a small part of the scientific method. After a general question is formulated, the scientific method consists of: designing an experiment, collecting data through observation and experimentation, testing hypotheses, and reporting overall conclusions. The conclusions themselves lead to other research ideas, making this process a continuous flow of adding to the body of knowledge about the phenomena being studied.



Others may choose a more formalized and detailed set of procedures, but the general concepts of inspiration, design, experimentation, and conclusion allow one to see the whole process.

This page titled [14.1: Procedures of Hypotheses Testing and the Scientific Method](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Maurice A. Geraghty](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- **10.1: Procedures of Hypotheses Testing and the Scientific Method** by [Maurice A. Geraghty](#) is licensed [CC BY-SA 4.0](#). Original source: <http://nebula2.deanza.edu/~mo/holisticInference.html>.

14.2: Formulate General Research Questions

Most general questions start with an inspiration or an idea about a topic or phenomenon of interest. Some examples of general questions:

- (Health Care) Would a public single payer health care system be more effective than the current private insurance system?
- (Labor) What is the effect of undocumented immigration and outsourcing of jobs on the current unemployment rate?
- (Economy) Is the federal economic stimulus package effective in lessening the impact of the recession?
- (Education) Are colleges too expensive for students today?

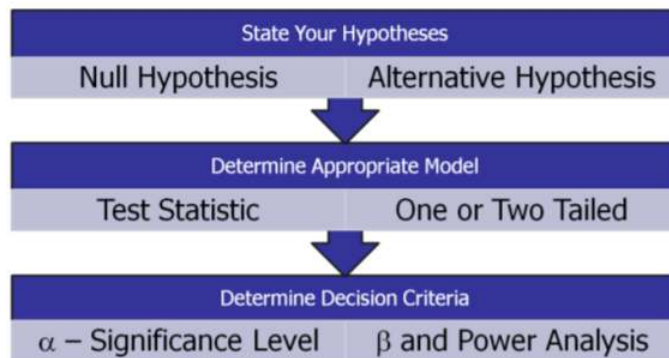
It is important to not be so specific in choosing these general questions. On the basis of available or potentially available data, we can decide later what specific research hypotheses will be formulated and tested to address the general question. During the data collection and testing process other ideas may come up and we may choose to redefine the general question. However, we always want to have an overriding purpose for our research.

This page titled [14.2: Formulate General Research Questions](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Maurice A. Geraghty](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [10.2: Formulate General Research Questions](#) by [Maurice A. Geraghty](#) is licensed [CC BY-SA 4.0](#). Original source: <http://nebula2.deanza.edu/~mo/holisticInference.html>.

14.3: Design Research Hypotheses and Experiment

After developing a general question and having some sense of the data that is available or that is collected, we then design an experiment and a set of **hypotheses**.



Hypotheses and Hypothesis

Testing For purposes of testing, we need to design hypotheses that are statements about population parameters. Some examples of hypotheses:

At least 20% of juvenile offenders are caught and sentenced to prison.

- The mean monthly income for college graduates is over \$5000.
- The mean standardized test score for schools in Cupertino is the same as the mean scores for schools in Los Altos.
- The lung cancer rates in California are lower than the rates in Texas.
- The standard deviation of the New York Stock Exchange today is greater than 10 percentage points per year.

These same hypotheses could be written in symbolic notation:

- $p \geq 0.20$
- $\mu > 5000$
- $\mu_1 = \mu_2$
- $p_1 < p_2$
- $\sigma > 10$

Hypothesis Testing is a procedure, based on sample evidence and probability theory, used to determine whether the hypothesis is a reasonable statement and should not be rejected, or is unreasonable and should be rejected. This hypothesis that is tested is called the **Null Hypothesis** and is designated by the symbol H_0 . If the Null Hypothesis is unreasonable and needs to be rejected, then the research supports an **Alternative Hypothesis** designated by the symbol H_a .

Definition: Null Hypothesis (H_0)

A statement about the value of a population parameter that is assumed to be true for the purpose of testing.

Definition: Alternative Hypothesis (H_a)

A statement about the value of a population parameter that is assumed to be true if the Null Hypothesis is rejected during testing.

From these definitions it is clear that the Alternative Hypothesis will necessarily contradict the Null Hypothesis; both cannot be true at the same time. Some other important points about hypotheses:

- Hypotheses must be statements about population parameters, never about sample statistics.
- In most hypotheses tests, equality ($=$, $\leq$, $\geq$) will be associated with the Null Hypothesis while non-equality ($\neq$, $<$, $>$) will be associated with the Alternative Hypothesis.

- It is the Null Hypothesis that is always tested in attempt to “disprove” it and support the Alternative Hypothesis. This process is analogous in concept to a “proof by contradiction” in Mathematics or Logic, but supporting a hypothesis with a level of confidence is not the same as an absolute mathematical proof.

Examples of Null and Alternative Hypotheses:

- $H_o : p \leq 0.20$ $H_a : p > 0.20$
- $H_o : \mu \leq 5000$ $H_a : \mu > 5000$
- $H_o : \mu_1 = \mu_2$ $H_a : \mu_1 \neq \mu_2$
- $H_o : p_1 \geq p_2$ $H_a : p_1 < p_2$
- $H_o : \sigma \leq 10$ $H_a : \sigma > 10$

Statistical Model and Test Statistic

To test a hypothesis we need to use a **statistical model** that describes the behavior for data and the type of population parameter being tested. Because of the Central Limit Theorem, many statistical models are from the Normal Family, most importantly the Z , t , χ^2 , and F distributions. Other models that are used when the Central Limit Theorem is not appropriate are called non-parametric Models and will not be discussed here.

Each chosen model has requirements of the data called **model assumptions** that should be checked for appropriateness. For example, many models require that the sample mean have approximately a Normal Distribution, something that may not be true for some smaller or heavily skewed data sets.

Once the model is chosen, we can then determine a **test statistic**, a value derived from the data that is used to decide whether to **reject** or **fail to reject** the Null Hypothesis.

Examples of Statistical Models and Test Statistics

Statistical Model	Test Statistic
Mean vs. Hypothesized Value	$t = \frac{\bar{X} - \mu_o}{s/\sqrt{n}}$
Proportion vs. Hypothesized Value	$Z = \frac{\hat{p} - p_o}{\sqrt{\frac{p_o(1-p_o)}{n}}}$
Variance vs. Hypothesized Value	$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$

Errors in Decision Making

Whenever we make a decision or support a position, there is always a chance we make the wrong choice. The hypothesis testing process requires us to either to reject the Null Hypothesis and support the Alternative Hypothesis or fail to reject the Null Hypothesis. This creates the possibility of two types of error:

- Type I Error Rejecting the null hypothesis when it is actually true.
- Type II Error Failing to reject the null hypothesis when it is actually false.

	Fail to Reject H_o	Reject H_o
H_o is true	Correct Decision	Type I error
H_o is False	Type II error	Correct Decision

In designing hypothesis tests, we need to carefully consider the probability of making either one of these errors.

Example: Pharmaceutical research

Recall the two news stories discussed earlier. In the first story, a drug company marketed a suppository that was later found to be ineffective (and often dangerous) in treatment. Before marketing the drug, the company determined that the drug was effective in treatment, meaning that the company rejected a Null Hypothesis that the suppository had no effect on the disease. This is an example of Type I error.

In the second story, research was abandoned when the testing showed Interferon was ineffective in treating a lung disease. The company in this case failed to reject a Null Hypothesis that the drug was ineffective. What if the drug really was effective? Did the company make Type II error? Possibly, but since the drug was never marketed, we have no way of knowing the truth.

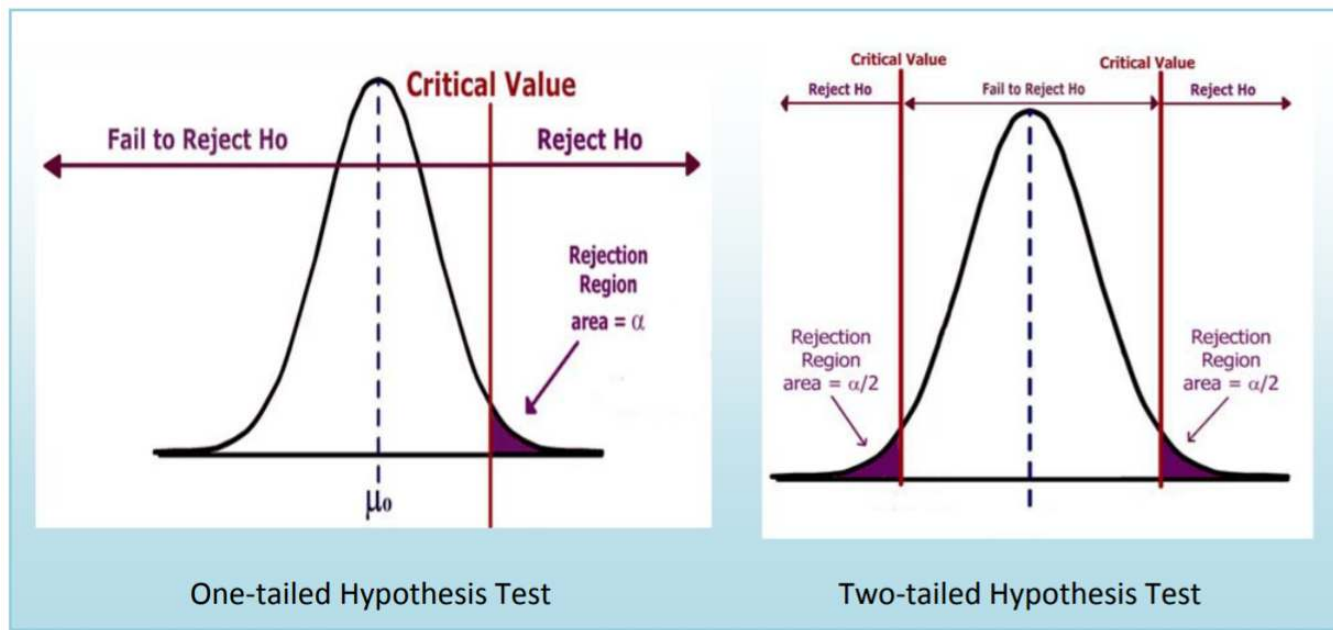
These stories highlight the problem of statistical research: errors can be analyzed using probability models, but there is often no way of identifying specific errors. For example, there are unknown innocent people in prison right now because a jury made Type I error in wrongfully convicting defendants. We must be open to the possibility of modification or rejection of currently accepted theories when new data is discovered.

In designing an experiment, we set a maximum probability of making Type I error. This probability is called the **level of significance** or **significance level** of the test and is designated by the Greek letter α , read as alpha. The analysis of Type II error is more problematic since there are many possible values that would satisfy the Alternative Hypothesis. For a specific value of the Alternative Hypothesis, the design probability of making Type II error is called **Beta (β)** which will be analyzed in detail later in this section.

Critical Value and Rejection Region

Once the significance level of the test is chosen, it is then possible to find the region(s) of the probability distribution function of the test statistic that would allow the Null Hypothesis to be rejected. This is called the **Rejection Region**, and the boundary between the Rejection Region and the “Fail to Reject” is called the **Critical Value**.

There can be more than one critical value and rejection region. What matters is that the total area of the rejection region equals the significance level α .



One and Two tailed Tests

A test is one-tailed when the Alternative Hypothesis, H_a , states a direction, such as:

H_a : The mean income of females is less than or equal to the mean income of males.

H_a : The mean income of females is greater than that of males.

Since equality is usually part of the Null Hypothesis, it is the Alternative Hypothesis which determines which tail to test.

A test is two-tailed when no direction is specified in the alternate hypothesis H_a , such as:

H_o : The mean income of females is equal to the mean income of males.

H_a : The mean income of females is not equal to the mean income of the males.

In a two tailed-test, the significance level is split into two parts since there are two rejection regions. In hypothesis testing, in which the statistical model is symmetrical (eg: the Standard Normal Z or Student's t distribution) these two regions would be equal. There is a relationship between a confidence interval and a two-tailed test: if the level of confidence for a confidence interval is equal to $1 - \alpha$, where α is the significance level of the two-tailed test, the critical values would be the same.

Here are some examples for testing the mean μ against a hypothesized value μ_0 :

Note

$H_a : \mu > \mu_0$ means test the upper tail and is also called a right-tailed test.

$H_a : \mu < \mu_0$ means test the lower tail and is also called a left-tailed test.

$H_a : \mu \neq \mu_0$ means test both tails.

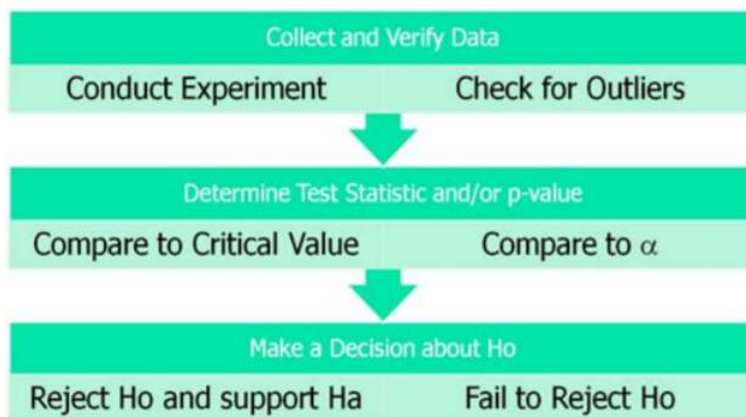
Deciding when to conduct a one or two-tailed test is often controversial and many authorities even go so far as to say that only two-tailed tests should be conducted. Ultimately, the decision depends on the wording of the problem. If we want to show that a new diet reduces weight, we would conduct a lower tailed test, since we don't care if the diet causes weight gain. If instead, we wanted to determine if mean crime rate in California was different from the mean crime rate in the United States, we would run a two-tailed test, since different implies greater than or less than.

This page titled [14.3: Design Research Hypotheses and Experiment](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Maurice A. Geraghty](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- **10.3: Design Research Hypotheses and Experiment** by [Maurice A. Geraghty](#) is licensed [CC BY-SA 4.0](#). Original source: <http://nebula2.deanza.edu/~mo/holisticInference.html>.

14.4: Collect and Analyze Experimental Data

After designing the experiment, we would then collect and verify the data. For the purposes of statistical analysis, we will assume that all sampling is either random or uses an alternative technique that adequately simulates a random sample.



Data Verification

After collecting the data but before running the test, we need to verify the data. First, get a picture of the data by making a graph (histogram, dot plot, box plot, etc). Check for skewness, shape and any potential outliers in the data.

Working with Outliers

An outlier is a data point that is far removed from the other entries in the data set. Outliers could be caused by:

- Mistakes made in recording data
- Data that don't belong in population
- True rare events

The first two cases are simple to deal with since we can correct errors or remove data that does not belong in the population. The third case is more problematic as extreme outliers will increase the standard deviation dramatically and heavily skew the data.

In *The Black Swan*, Nicholas Taleb argues that some populations with extreme outliers should not be analyzed with traditional confidence intervals and hypothesis testing.⁷² He defines a Black Swan to be an unpredictable extreme outlier that causes dramatic effects on the population. A recent example of a Black Swan was the catastrophic drop in the value of unregulated Credit Default Swap (CDS) real estate insurance investments causing the near collapse of the international banking system in 2008. The traditional statistical analysis that measured the risk of the CDS investments did not take into account the consequence of a rapid increase in the number of foreclosures of homes. In this case, statistics that measure investment performance and risk were useless and created a false sense of security for large banks and insurance companies.

Example: Realtor home sales

Here are the quarterly home sales for 10 realtors

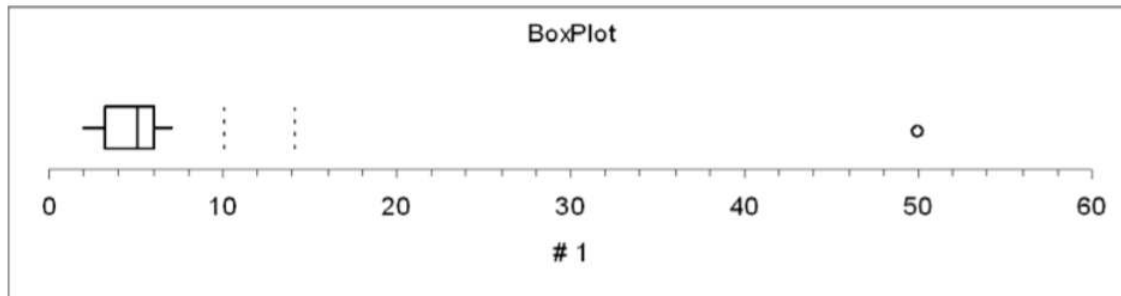
2 2 3 4 5 5 6 6 7 50

	With outlier	Without outlier
Mean	9.00	4.44
Median	5.00	5.00
Standard Deviation	14.51	1.81
Interquartile Range	3.00	3.50

In this example, the number 50 is an outlier. When calculating summary statistics, we can see that the mean and standard deviation are dramatically affected by the outlier, while the median and the interquartile range (which are based on the ranking of the data) are hardly changed. One solution when dealing with a population with extreme outliers is to use inferential statistics using the ranks of the data, also called non-parametric statistics.

Using Box Plot to find outliers

- The “box” is the region between the 1st and 3rd quartiles.
- Possible outliers are more than 1.5 IQR’s from the box (inner fence)
- Probable outliers are more than 3 IQR’s from the box (outer fence)
- In the box plot below, which illustrates the realtor example, the dotted lines represent the “fences” that are 1.5 and 3 IQR’s from the box. See how the data point 50 is well outside the outer fence and therefore an almost certain outlier.



The Logic of Hypothesis Testing

After the data is verified, we want to conduct the hypothesis test and come up with a decision: whether or not to reject the Null Hypothesis. The decision process is similar to a “proof by contradiction” used in mathematics:

- We assume H_0 is true before observing data and design H_a to be the complement of H_0 .
- Observe the data (evidence). How unusual are these data under H_0 ?
- If the data are too unusual, we have “proven” H_0 is false: reject H_0 and support H_a (strong statement).
- If the data are not too unusual, we fail to reject H_0 . This “proves” nothing and we say data are inconclusive. (weak statement) .
- We can never “prove” H_0 , only “disprove” it.
- “Prove” in statistics means support with $(1 - \alpha)100\%$ certainty. (example: if $\alpha = .05$, then we are at least 95% confident in our decision to reject H_0).

Decision Rule – Two methods, Same Decision

Earlier we introduced the idea of a **test statistic** which is a value calculated from the data under the appropriate Statistical Model from the data that can be compared to the **critical value** of the Hypothesis test. If the test statistic falls in the **rejection region** of the statistical model, we reject the Null Hypothesis.

Recall that the critical value was determined by design on the basis of the chosen **level of significance α** . The more preferred method of making decisions is to calculate the probability of getting a result as extreme as the value of the test statistic. This probability is called the **p-value**, and can be compared directly to the significance level.

Definition: p-value

p-value: the probability, assuming that the null hypothesis is true, of getting a value of the test statistic at least as extreme as the computed value for the test.

- If the *p*-value is smaller than the significance level α , H_0 is rejected.
- If the *p*-value is larger than the significance level α , H_0 is not rejected.

Comparing p -value to α

Both the p -value and α are probabilities of getting results as extreme as the data assuming H_o is true.

The p -value is determined **by the data** and is related to the actual probability of making Type I error (rejecting a true Null Hypothesis). The smaller the p -value, the smaller the chance of making Type I error and therefore, the more likely we are to reject the Null Hypothesis.

The significance level α is determined **by the design** and is the maximum probability we are willing to accept of rejecting a true H_o .

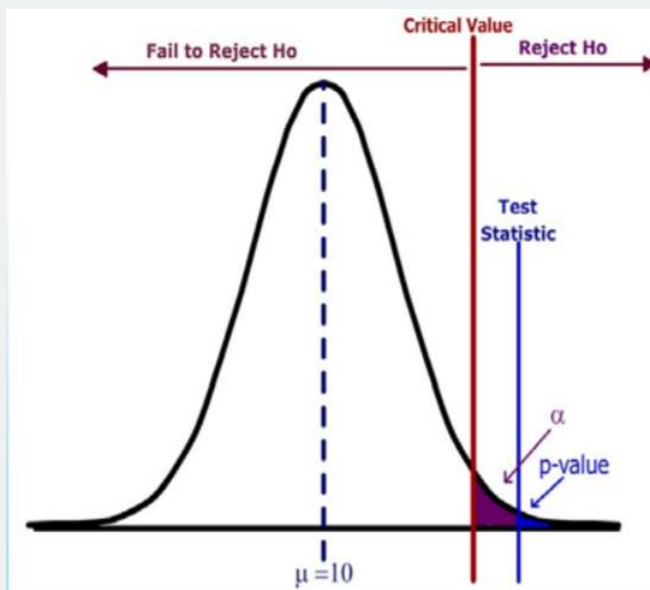
📌 Two Decision Rules lead to the same decision.

1. If the test statistic lies in the rejection region, reject H_o . (critical value method)
2. If the p -value $< \alpha$, reject H_o . (p -value method)

This p -value method of comparison is preferred to the critical value method because the rule is the same for all statistical models: Reject H_o if p -value $< \alpha$.

Let's see why these two rules are equivalent by analyzing a test of mean vs. hypothesized value.

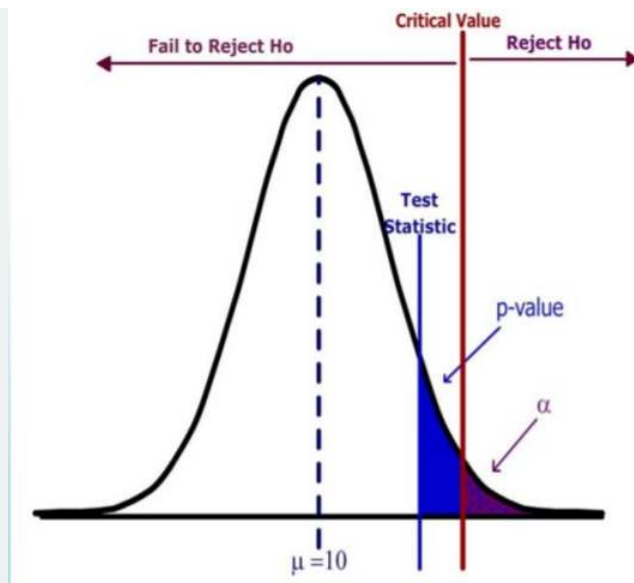
📌 Decision is Reject H_o



- $H_o : \mu = 10$ $H_a : \mu > 10$
- Design: Critical value is determined by significance level α .
- Data Analysis: p -value is determined by test statistic
- Test statistic falls in rejection region.
- p -value (blue) $< \alpha$ (purple)
- Reject H_o .
- Strong statement: Data supports the Alternative Hypothesis.

In this example, the test statistic lies in the rejection region (the area to the right of the critical value). The p -value (the area to the right of the test statistic) is less than the significance level (the area to the right of the critical value). The decision is to Reject H_o .

📌 Decision is Fail to Reject H_o



- $H_o : \mu = 10$ $H_a : \mu > 10$
- Design: critical value is determined by significance level α .
- Data Analysis: p -value is determined by test statistic
- Test statistic does not fall in the rejection region.
- p -value (blue) $>$ α (purple)
- Fail to Reject H_o .
- Weak statement: Data is inconclusive and does not support the Alternative Hypothesis.

In this example, the Test Statistic does not lie in the Rejection Region. The p -value (the area to the right of the test statistic) is greater than the significance level (the area to the right of the critical value). The decision is Fail to Reject H_o .

This page titled [14.4: Collect and Analyze Experimental Data](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Maurice A. Geraghty](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [10.4: Collect and Analyze Experimental Data](#) by [Maurice A. Geraghty](#) is licensed [CC BY-SA 4.0](#). Original source: <http://nebula2.deanza.edu/~mo/holisticInference.html>.

14.5: Report Conclusions in Non-statistical Language

The hypothesis test has been conducted and we have reached a decision. We must now communicate these conclusions so they are complete, accurate, and understood by the targeted audience. How a conclusion is written is open to subjective analysis, but here are a few suggestions.

Be consistent with the results of the Hypothesis Test

Rejecting H_o requires a **strong statement** in support of H_a , while failing to reject H_o does NOT support H_o , but requires a **weak statement** of insufficient evidence to support H_a .

Example



A researcher wants to support the claim that, on average, students send more than 1000 text messages per month, and the research hypotheses are $H_o : \mu = 1000$ vs. $H_a : \mu > 1000$

Conclusion if H_o is rejected: The mean number of text messages sent by students exceeds 1000.

Conclusion if H_o is not rejected: There is insufficient evidence to support the claim that the mean number of text messages sent by students exceeds 1000.

Use language that is clearly understood in the context of the problem

Do not use technical language or jargon, but instead refer back to the language of the original general question or research hypotheses. Saying less is better than saying more.

Example

A test supported the Alternative Hypothesis that housing prices and size of homes in square feet were positively correlated. Compare these two conclusions and decide which is clearer:



Solution

Conclusion 1: By rejecting the Null Hypothesis we are inferring that the Alternative Hypothesis is supported and that there exists a significant correlation between the independent and dependent variables in the original problem comparing home prices to square footage.

Conclusion 2: Homes with more square footage generally have higher prices.

Limit the inference to the population that was sampled

Care must be taken to describe the population being sampled and understand that the any claim is limited to this sampled population. If a survey was taken of a subgroup of a population, then the inference applies only to the subgroup.

For example, studies by pharmaceutical companies will only test adult patients, making it difficult to determine effective dosage and side effects for children. “In the absence of data, doctors use their medical judgment to decide on a particular drug and dose for children. ‘Some doctors stay away from drugs, which could deny needed treatment,’ Blumer says. ‘Generally, we take our best guess based on what’s been done before.’ The antibiotic chloramphenicol was widely used in adults to treat infections resistant to penicillin. But many newborn babies died after receiving the drug because their immature livers couldn’t break down the antibiotic.”⁷³ We can see in this example that applying inference of the drug testing results on adults to the un-sampled children led to tragic results.

Report sampling methods that could question the integrity of the random sample assumption

In practice it is nearly impossible to choose a random sample, and scientific sampling techniques that attempt to simulate a random sample need to be checked for bias caused by under-sampling.

Telephone polling was found to under-sample young people during the 2008 presidential campaign because of the increase in cell phone only households. Since young people were more likely to favor Obama, this caused bias in the polling numbers. Additionally, caller ID has dramatically reduced the percentage of successful connections to people being surveyed. The pollster Jay Leve of SurveyUSA said telephone polling was “doomed” and said his company was already developing new methods for polling.⁷⁴

Sampling that didn’t occur over the weekend may exclude many full time workers while self-selected and unverified polls (such as ratemyprofessors.com) could contain immeasurable bias.

Conclusions should address the potential or necessity of further research, sending the process back to the first procedure

Answers often lead to new questions. If changes are recommended in a researcher’s conclusion, then further research is usually needed to analyze the impact and effectiveness of the implemented changes. There may have been limitations in the original research project (such as funding resources, sampling techniques, unavailability of data) that warrant more comprehensive studies.

For example, a math department modifies its curriculum based on the improved student success rates of an experimental course. The department would want to do further study of student outcomes to assess the effectiveness of the new program.

This page titled [14.5: Report Conclusions in Non-statistical Language](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Maurice A. Geraghty](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- **10.5: Report Conclusions in Non-statistical Language** by [Maurice A. Geraghty](#) is licensed [CC BY-SA 4.0](#). Original source: <http://nebula2.deanza.edu/~mo/holisticInference.html>.

14.6: Test of Mean vs. Hypothesized Value – A Complete Example

Example – Soy sauce production



A food company has a policy that the stated contents of a product match the actual results. A **General Question** might be “Does the stated net weight of a food product match the actual weight?” The quality control statistician decides to test the 16 ounce bottle of Soy Sauce and must now **design the experiment**.

The quality-control statistician has been given the authority to sample 36 bottles of soy sauce and knows from past testing that the population standard deviation is 0.5 ounces. The model will be a **test of population mean vs. hypothesized value** of 16 oz. A two-tailed test is selected since the company is concerned about both overfilling and underfilling the bottles as the stated policy is that the stated weight should match the actual weight of the product.

Research Hypotheses:

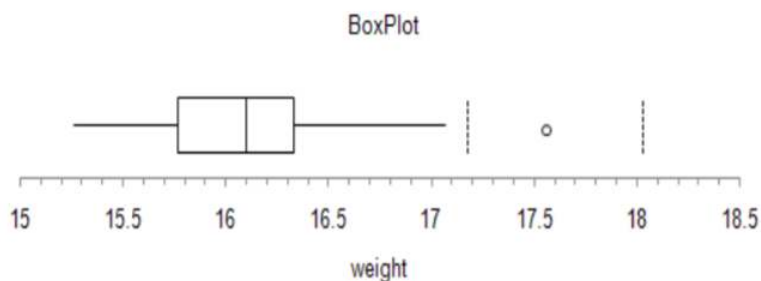
$H_o : \mu = 16$ (The filling machine is operating properly)

$H_a : \mu \neq 16$ (The filling machine is not operating properly)

Since the population standard deviation is known the **test statistic** will be $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$. This model is appropriate since the sample size assures that the distribution of the sample mean is approximately Normal due to the Central Limit Theorem.

Type I error would be to reject the Null Hypothesis and say that the machine is not running properly when in fact it was operating properly. Since the company does not want to needlessly stop production and recalibrate the machine, the statistician chooses to limit the probability of Type I error by setting the **level of significance** (α) to 5%.

The statistician now conducts the experiment and samples 36 bottles over one hour and determines from a box plot of the data that there is one unusual observation of 17.56 ounces. The value is rechecked and kept in the data set.

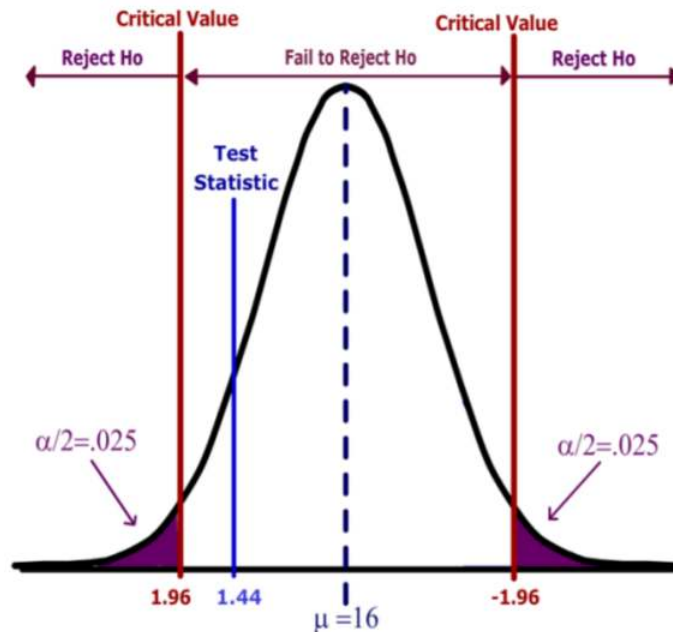


Next, the sample mean and the test statistic are calculated.

$$\bar{X} = 16.12 \text{ ounces} \quad Z = \frac{16.12 - 16}{0.5/\sqrt{36}} = 1.44$$

The **decision rule** under the critical value method would be to reject the Null Hypothesis when the value of the test statistic is in the rejection region. In other words, reject H_o when $Z > 1.96$ or $Z < -1.96$.

Based on this result, the decision is **fail to reject H_o** , since the test statistic does not fall in the rejection region.

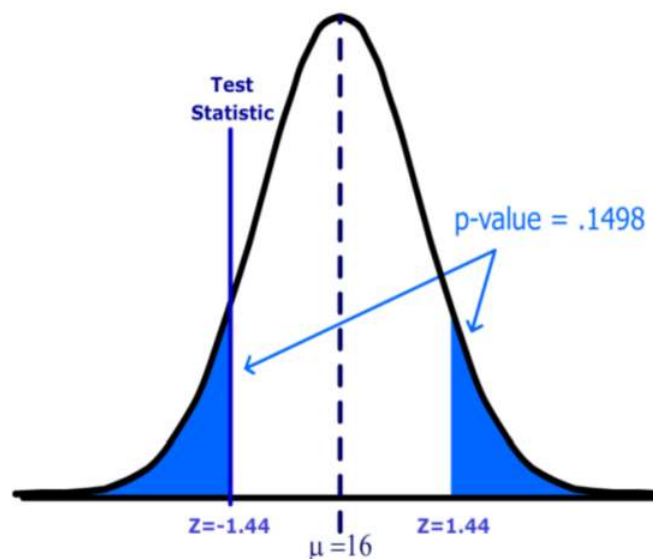


Alternatively (and preferably) the statistician could use the p -value method of decision rule. The p -value for a two-tailed test must include all values (positive and negative) more extreme than the Test Statistic, so in this example we find the probability that $Z < -1.44$ or $Z > 1.44$ (the area shaded blue).

Using a calculator, computer software or a Standard Normal table, the **p -value=0.1498**. Since the p -value is greater than α the decision again is **fail to reject H_o** .

Finally the statistician must **report the conclusions** and make a recommendation to the company's management:

“There is insufficient evidence to conclude that the machine that fills 16 ounce soy sauce bottles is operating improperly. This conclusion is based on 36 measurements taken during a single hour's production run. I recommend continued monitoring of the machine during different employee shifts to account for the possibility of potential human error”.



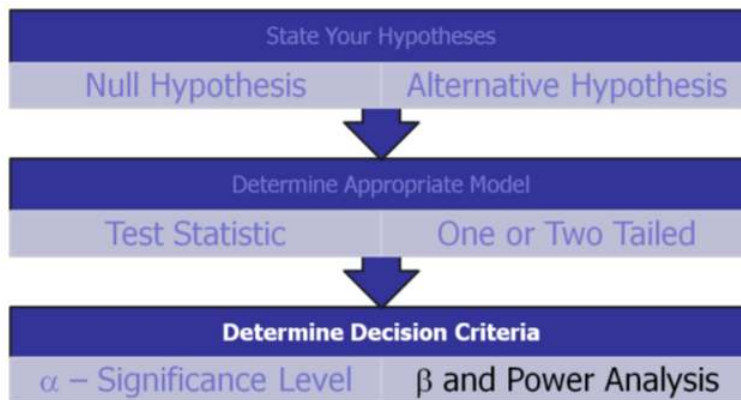
The statistician makes the weak statement and is not stating that the machine is running properly, only that there is not enough evidence to state that the machine is running improperly. The statistician also reports concerns about the sampling of only one shift of employees (restricting the inference to the sampled population) and recommends repeating the experiment over several shifts.

This page titled [14.6: Test of Mean vs. Hypothesized Value – A Complete Example](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Maurice A. Geraghty](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- **10.6: Test of Mean vs. Hypothesized Value – A Complete Example** by [Maurice A. Geraghty](#) is licensed [CC BY-SA 4.0](#). Original source: <http://nebula2.deanza.edu/~mo/holisticInference.html>.

14.7: Type II Error and Statistical Power

In the prior example, the statistician failed to reject the Null Hypothesis because the probability of making Type I error (rejecting a true Null Hypothesis) exceeded the significance level of 5%. However, the statistician could have made Type II error if the machine is really operating improperly. One of the important and often overlooked tasks is to analyze the probability of making Type II error (β). Usually statisticians look at statistical power which is the complement of β .



Beta (β): The probability of failing to reject the null hypothesis when it is actually false.

Power (or Statistical Power): The probability of rejecting the null hypothesis when it is actually false.

Both beta and power are calculated for specific possible values of the Alternative Hypothesis.

	Fail to Reject H_o	Reject H_o
H_o is true	$1 - \alpha$	α Type I error
H_o is false	β Type II error	$1 - \beta$ Power

If a hypothesis test has low power, then it would be difficult to reject H_o , even if H_o were false; the research would be a waste of time and money. However, analyzing power is difficult in that there are many values of the population parameter that support H_a . For example, in the soy sauce bottling example, the Alternative Hypothesis was that the mean was not 16 ounces. This means the machine could be filling the bottles with a mean of 16.0001 ounces, making H_a technically true. So when analyzing power and Type II error, we need to choose a value for the population mean under the **Alternative Hypothesis (μ_a)** that is “**practically different**” from the **mean under the Null Hypothesis (μ_o)**. This practical difference is called the **effect size**.

Definition: Effect size

Effect Size: The “practical difference” between μ_o and $\mu_a = |\mu_o - \mu_a|$

where

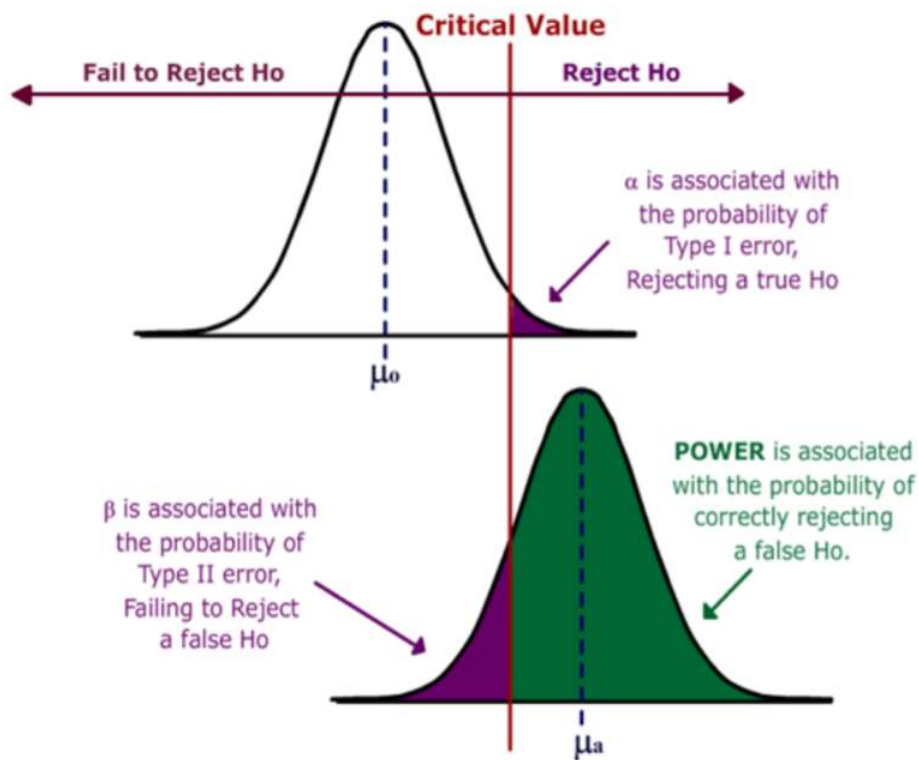
μ_o : The value of the population mean under the Null Hypothesis

μ_a : The value of the population mean under the Alternative Hypothesis

Suppose we are conducting a one-tailed test of the population mean:

$$H_o : \mu = \mu_0 \quad H_a : \mu > \mu_0$$

Consider the two graphs shown below. The top graph is the distribution of the sample mean under the Null Hypothesis, which was covered in an earlier section. The area to the right of the critical value is the rejection region.



We now add the bottom graph, which represents the distribution of the sample mean under the Alternative Hypothesis for the specific value μ_a .

We can now measure the Power of the test (the area in green) and beta (the area in purple) on the lower graph.

There are several methods of increasing Power, but they all have trade-offs:

Ways to Increase Power	Trade off
Increase Sample Size	Increased cost or unavailability of data
Increase Significance level (α)	More like to Reject a true H_o (Type I error)
Choose a value of μ_a further from μ_o	Result may be less meaningful
Redefine population to lower standard deviation	Result may be too limited to have value
Conduct as a one-tail rather than a two-tail test	May produce a biased result

Example: Bus brake pads

Bus brake pads are claimed to last on average at least 60,000 miles and the company wants to test this claim. The bus company considers a “practical” value for purposes of bus safety to be that the pads last at least 58,000 miles. If the standard deviation is 5,000 and the sample size is 50, find the power of the test when the mean is really 58,000 miles. (Assume $\alpha = .05$)



Solution

First, find the critical value of the test.

Reject H_o when $Z < -1.645$

Next, find the value of that corresponds to the critical value.

$$\bar{X} = \mu_o + \frac{Z\sigma}{\sqrt{n}} = 60000 - (1.645)(5000)/\sqrt{50} = 58837$$

H_o is rejected when $\bar{X} < 58837$

Finally, find the probability of rejecting H_o if H_a is true.

$$\begin{aligned} P(\bar{X} < 58837) &= P\left(Z < \frac{(58837 - \mu_a)}{\sigma/\sqrt{n}}\right) \\ &= P\left(Z < \frac{(58837 - 58000)}{5000/\sqrt{50}}\right) \\ &= P(Z < 1.18) \\ &= .8810 \end{aligned}$$

Therefore, this test has 88% power and β would be 12%

Power Calculation Values

Input Values

$\mu_o = 60,000$ miles

$\mu_a = 58,000$ miles

$\alpha = 0.05^*$

$n = 50$

$\sigma = 5000$ miles

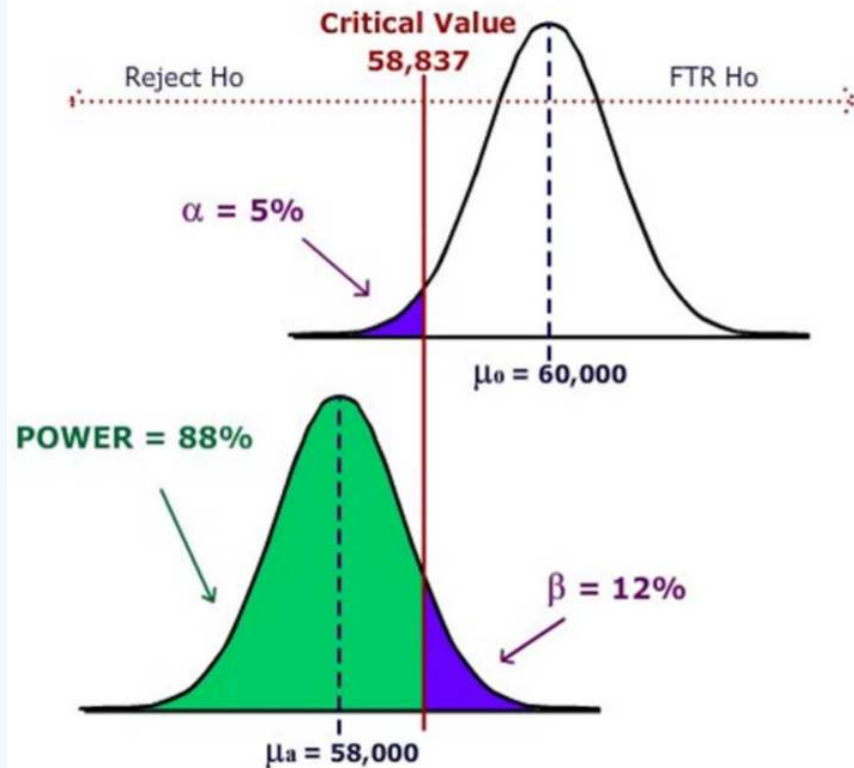
Calculated Values

Effect Size = 2000 miles

Critical Value = 58,837 miles

$\beta = 0.1190$ or about 12%

Power = 0.8810 or about 88%



This page titled [14.7: Type II Error and Statistical Power](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Maurice A. Geraghty](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [10.7: Type II Error and Statistical Power](#) by [Maurice A. Geraghty](#) is licensed [CC BY-SA 4.0](#). Original source: <http://nebula2.deanza.edu/~mo/holisticInference.html>.

14.8: New Models for One Population Inference, Similar Procedures

The procedures outlined for the test of population mean vs. hypothesized value with known population standard deviation will apply to other models as well. All that really changes is the test statistic.

Examples of some other one population models:

- Test of population mean vs. hypothesized value, population standard deviation unknown.
- Test of population proportion vs. hypothesized value.
- Test of population standard deviation (or variance) vs. hypothesized value.

Test of population mean with unknown population standard deviation

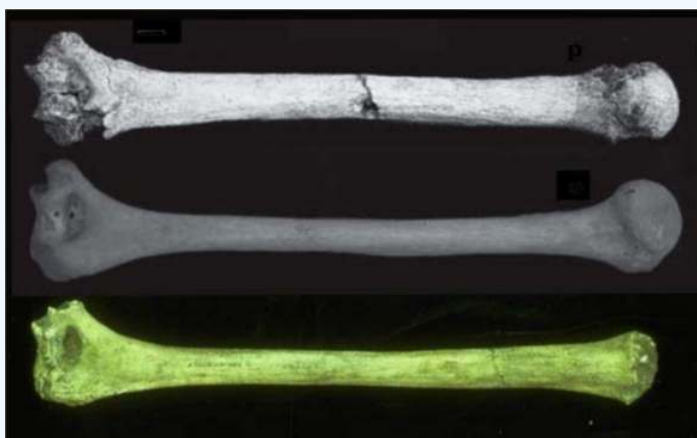
The test statistic for the one sample case changes to a Student's t distribution with degrees of freedom equal to

$$n - 1 : t = \frac{\bar{X} - \mu_o}{s/\sqrt{n}}$$

The shape of the t distribution is similar to the Z , except for the fact that the tails are fatter, so the logic of the decision rule is the same as for the Z test statistic.

Example: Archaeology

Humerus bones from the same species have approximately the same length-to-width ratios. When fossils of humerus bones are discovered, archaeologists can determine the species by examining this ratio. It is known that Species A has a mean ratio of 9.6. A similar Species B has a mean ratio of 9.1 and is often confused with Species A. 21 humerus bones were unearthed in an area that was originally thought to be inhabited Species A. (Assume all unearthed bones are from the same species.)



- Design a test in which the alternative hypothesis would be the humerus bones were not from Species A.
- Determine the power of this test if the bones actually came from Species B (assume a standard deviation of 0.7)
- Conduct the test using at a 5% significance level and state overall conclusions.

Solution

- Research Hypotheses

$H_o : \mu = 9.6$ (The humerus bones are from Species A)

$H_a : \mu \neq 9.6$ (The humerus bones are not from Species A)

Significance level: $\alpha = .05$

Test Statistic (Model): t -test of mean vs. hypothesized value, unknown standard deviation

Model Assumptions: we may need to check the data for extreme skewness as the distribution of the sample mean is assumed to be approximately the Normal Distribution.

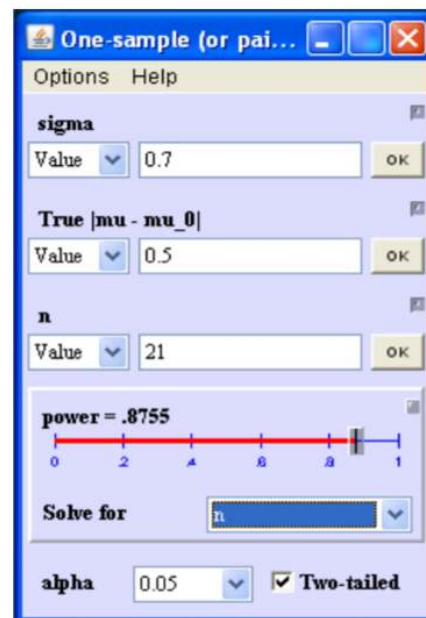
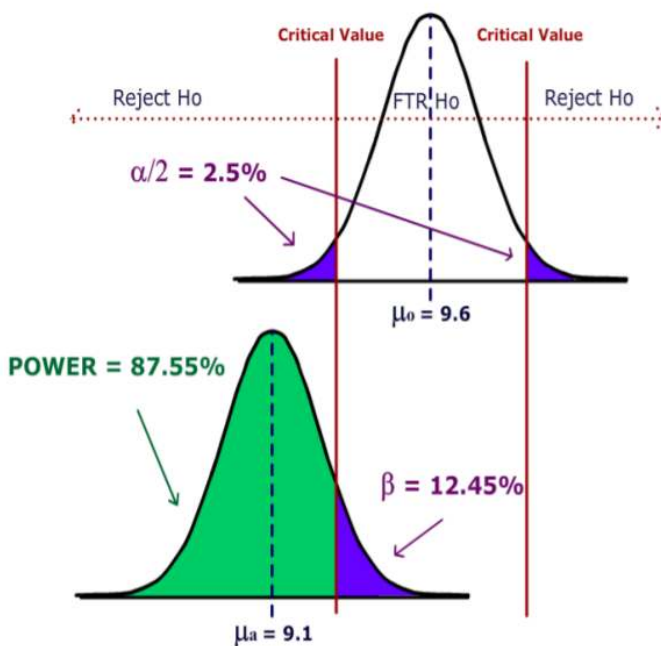
-

Information needed for Power Calculation

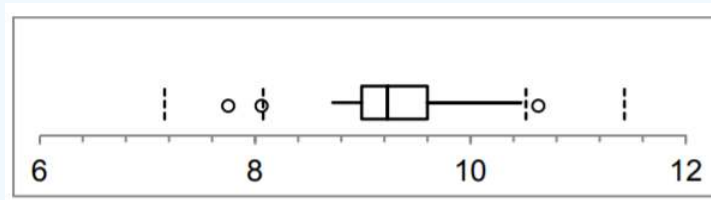
- $\mu_o = 9.6$ (Species A)
- $\mu_a = 9.1$ (Species B)
- Effect Size = $|m_o - m_a| = 0.5$
- $s = 0.7$ (given)
- $\alpha = .05$
- $n = 21$ (sample size)
- Two tailed test

Results using Online Power Calculator⁷⁵

- Power = .8755
- $\beta = 1 - \text{Power} = .1245$
- If humerus bones are from Species B, test has an 87.55% chance of correctly rejecting H_o and a maximum Type II error of 12.55%



c.



From MegaStat⁷⁶, $p\text{-value} = .0308$ and $\alpha = .05$.

Since $p\text{-value} < \alpha$, H_o is rejected and we support H_a .

Hypothesis Test: Mean vs. Hypothesized Value

9.60000	hypothesized value
9.26190	mean Data
0.66700	std. dev.
0.14555	std. error
21	n
20	df
-2.32	t
.0308	p-value (two-tailed)

Conclusion: The evidence supports the claim ($p\text{-value} < .05$) that the humerus bones are not from Species A. The small sample size limited the power of the test, which prevented us from making a more definitive conclusion. Further testing is recommended to determine if bones are from Species B or other unknown species.

We are also assuming that since the bones were unearthed in the same location, they came from the same species.

Test of population proportion vs. hypothesized value

When our data is categorical and there are only two possible choices (for example a yes/no question on a poll), we may want to make a claim about a proportion or a percentage of the population (p) being compared to a particular value (p_o). We will then use the sample proportion ($\hat{p}$) to test the claim.

📌 Test of proportion vs. hypothesized value

p = population proportion

p_o = population proportion under H_o

$\hat{p}$ = sample proportion

p_a = population proportion under H_a

Test Statistic:
$$Z = \frac{\hat{p} - p_o}{\sqrt{\frac{p_o(1-p_o)}{n}}}$$

Requirement for Normality Assumption: $np(1-p) > 5$

Example: Charity solicitation

In the past, 15% of the mail order solicitations for a certain charity resulted in a financial contribution. A new solicitation letter has been drafted and will be sent to a random sample of potential donors. A hypothesis test will be run to determine if the new letter is more effective. Determine the sample so that (1) the test will be run at the 5% significance level and (2) if the letter has an 18% success rate, (an effect size of 3%), the power of the test will be 95%. After determining the sample size, conduct the test.



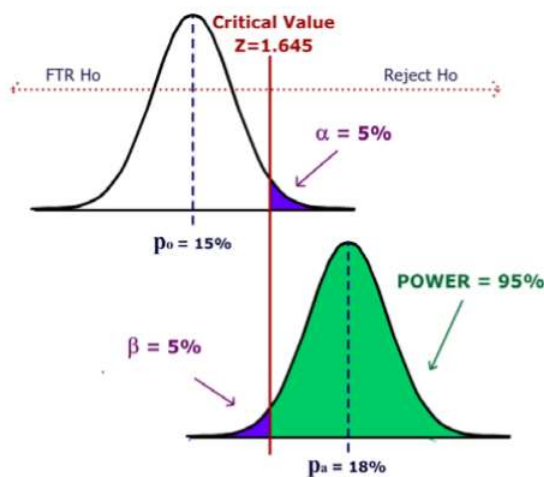
Solution

$H_o : p \leq 0.15$ (The new letter is not more effective.)

$H_a : p > 0.15$ (The new letter is more effective.)

Test Statistic – Z-test of proportion vs. hypothesized value

Information needed for Power Calculation	Results using Online Power Calculator ⁷⁵
• $p_o = 0.15$ (current letter) • $p_a = 0.18$ (potential new letter) • Effect Size = $p_a - p_o = 0.03$ • Desired Power = 0.95 • $\alpha = .05$ • One tailed test	• Sample size = 1652 • The charity sent out 1652 new solicitation letters to potential donors and ran the test, receiving 286 positive responses. • p-value for test = 0.0042



Sample size for one proportion

Options Help

Null value (p_0) = .15

Actual value (p) = .18

Sample size

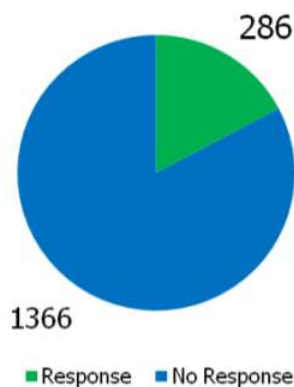
Value 1652 OK

Alternative $p > p_0$ Alpha 0.05

Method Normal approx

Power

Value 95 OK



Hypothesis test for proportion vs hypothesized value

Observed	Hypothesized	
0.1731	0.15	p (as decimal)
286/1652	248/1652	p (as fraction)
286.	247.8	X
1652	1652	n
	0.0088	std. error
	2.63	z
	.0042	p-value (one-tailed, upper)

Since $p\text{-value} < \alpha$, reject H_0 and support H_a . Since the $p\text{-value}$ is actually less than 0.01, we would go further and say that the data supports rejecting H_0 for $\alpha = .01$.

Conclusion: The evidence supports the claim that the new letter is more effective. The 1652 test letters were selected as a random sample from the charity's mailing list. All letters were sent at the same time period. The letters needed to be sent in a specific time period, so we were not able to control for seasonal or economic factors. We recommend testing both solicitation methods over the entire year to eliminate seasonal effects and to create a control group.

Test of population standard deviation (or variance) vs. hypothesized value

We often want to make a claim about the variability, volatility or consistency of a population random variable. Hypothesized values for the population variance (σ^2) or the standard deviation (σ) are tested with the Chi-square (χ^2) distribution.

Examples of Hypotheses:

- $H_0 : \sigma = 10$ $H_a : \sigma \neq 10$
- $H_0 : \sigma^2 = 100$ $H_a : \sigma^2 > 100$

The sample variance s^2 is used in calculating the Chi-square Test Statistic.

Test of variance vs. hypothesized value

σ^2 = population variance

σ_o^2 = population variance under H_0

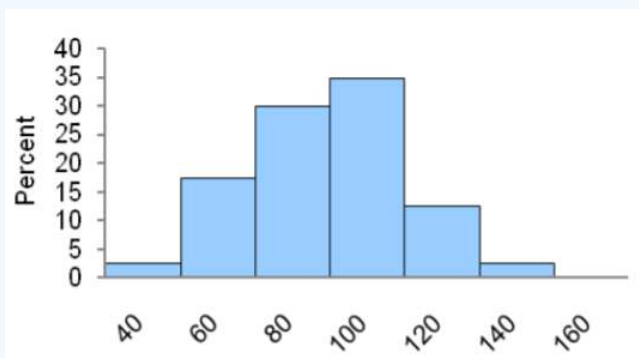
s^2 = sample variance

Test Statistic: $\chi^2 = \frac{(n-1)s^2}{\sigma_o^2}$

$n - 1$ = degrees of freedom

Example: Standardized testing

A state school administrator claims that the standard deviation of test scores for 8th grade students who took a life-science assessment test is less than 30, meaning the results for the class show consistency. An auditor wants to support that claim by analyzing 41 students' recent test scores. The test will be run at 1% significance level.



57	75	86	92	101	108	110	120	155
63	77	88	96	102	108	111	122	
66	78	88	96	107	109	115	135	
68	81	92	98	107	109	115	137	
72	82	92	99	107	110	118	139	

Solution

Design:

Research Hypotheses:

H_o : Standard deviation for test scores equals 30.

H_a : Standard deviation for test scores is less than 30.

Hypotheses In terms of the population variance:

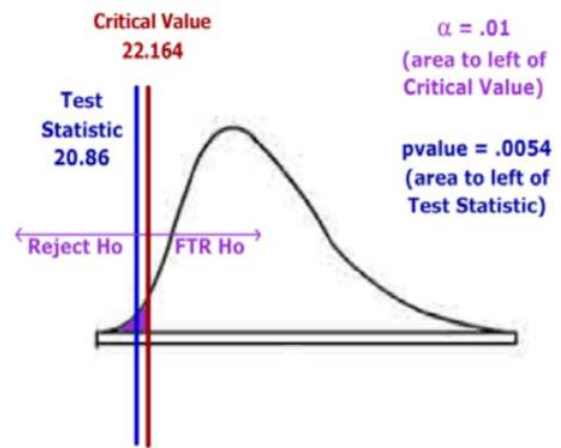
$H_o : \sigma^2 = 900$

$H_a : \sigma^2 < 900$

Results:

Chi-square Variance Test

900.000 hypothesized variance
 469.426 observed variance of Data
 41 n
 40 df
 20.86 chi-square
 .0054 p-value (one-tailed, lower)



Decision: Reject H_0

Conclusion: The evidence supports the claim ($p\text{-value} < .01$) that the standard deviation for 8th grade test scores is less than 30. The 40 test scores were the results of the recently administered exam to the 8th grade students. Since the exams were for the current class only, there is no assurance that future classes will achieve similar results. Further research would be to compare results to other schools that administered the same exam and to continue to analyze future class exams to see if the claim is holding true.

This page titled [14.8: New Models for One Population Inference, Similar Procedures](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Maurice A. Geraghty](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [10.8: New Models for One Population Inference, Similar Procedures](#) by [Maurice A. Geraghty](#) is licensed [CC BY-SA 4.0](#). Original source: <http://nebula2.deanza.edu/~mo/holisticInference.html>.

14.9: The p-value- Misconceptions and Proper Usage

One of the most misinterpreted concepts in Statistics is the p-value. In government studies and scientific research, there have been invalid conclusion based on misinterpreting the p -value. On March 7, 2016, in an unprecedented statement, the American Statistical Association released a paper, "Statement on Statistical Significance and P-Values", which offered principles to improve the conduct and interpretation of quantitative science.⁷⁷

The paper introduced 6 standards, which we will review individually.

1. P -values can indicate how incompatible the data are with a specified statistical model.

The p -value is the probability of getting data this extreme given H_o is true. This is a conditional probability and can be written as:

$$p\text{-value} = P(\text{getting this data or more extreme data} \mid H_o \text{ is true})$$

Example: Financial aid

A researcher wanted to show that the percentage of students at community colleges who receive financial aid exceeds 40%.

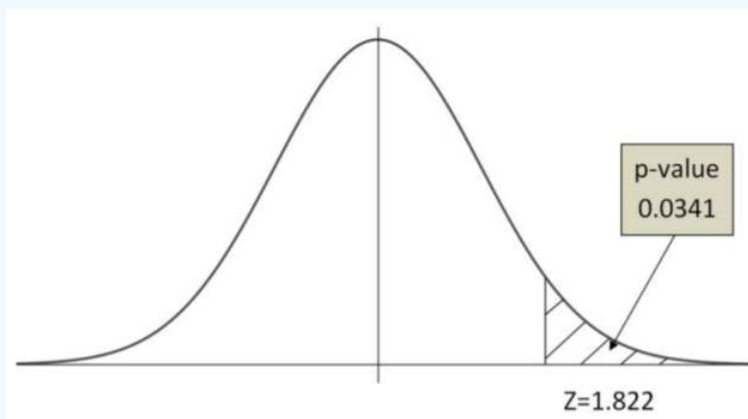
Solution

$H_o : p = 0.40$ (The proportion of community college students receiving financial aid is 0.40.).

$H_a : p > 0.40$ (The proportion of community college students receiving financial aid is over 0.40.).

The research sampled 874 students and found that 376 of them received financial aid. This works to a sample proportion $\hat{p} = 0.430$, which leads to a Z value of 1.822, if $p = 0.40$.

$$\begin{aligned} p\text{-value} &= P(\hat{p} > 0.430 \mid H_o \text{ is true}) \\ &= P(Z > 1.822) \\ &= 0.034 \end{aligned}$$



The probability of getting this sample proportion, or something larger given the actual proportion is 0.40 is equal 0.034.

2. P -values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.

After conducting an experiment, researchers would love to be able know the probability that their claim is true. Unfortunately, this probability cannot be calculated from the p -value alone.

Example: Financial aid

Let's return to the researcher who wanted to show that the percentage of students at community colleges who receive financial aid exceeds 40%. After conducting the research, the p -value was 0.034. Suppose the researcher wrote this conclusion:

"With 96.6% confidence, we conclude that the percentage of community college students who receive financial aid exceeds 40%".

This conclusion is invalid, and conclusions written with a similar misinterpretation has shown up in many published works. Let's explore the problem here.

The researcher is claiming that the probability that the alternative hypothesis is true is the complement of the p -value. In other words, the researcher is claiming the p -value is the probability H_o is true given this data. This researcher has flipped the conditionality in the p -value definition!

Researcher's misinterpretation: $p\text{-value} = P(H_o \text{ is true} \mid \text{Data this Extreme})$

Correct interpretation of p -value = $P(\text{Getting Data this Extreme} \mid H_o \text{ is true})$

In Chapter 5 on probability, we explored why $P(A|B)$ is not the same as $P(B|A)$.

Recall the testing for HIV example from Chapter 5

$$P(\text{Tests} + \mid \text{HIV-}) = 1350/9000 = 85\%$$

$$P(\text{HIV+} \mid \text{Tests+}) = 950/2300 = 41.3\%$$

Even though the test has a true-positive rate of 85%, there is only a 41% chance that someone who tests positive has HIV.

	HIV+ A	HIV- A'	Total
Test+ B	950	1350	2300
Test- B'	50	7650	7700
Total	1000	9000	10000

3. Scientific conclusions and business or policy decisions should not be based only on whether a p -value passes a specific threshold.

In any statistics course, we learn that having a p -value less than the significance level is evidence supporting the Alternative Hypothesis. This does not necessarily mean H_a is true or even probably true. There needs to be other reasoning as to why H_a might be true.

Some research journals, like Basic and Applied Social Psychology, now require research show “strong descriptive statistics, including effect sizes.”⁷⁸

Example: Financial aid

We will again return to financial aid example. After conducting the research, the p -value was 0.031. If we started with a significance level of 5%, the decision would be to Reject H_o and support the claim that the percentage of students at community colleges who receive financial aid exceeds 40%. However, if we started with a significance level of 1%, the decision would be to Fail to Reject H_o and there would not be enough evidence to support the claim that the percentage of students at community colleges who receive financial aid exceeds 40%. Even if H_o is rejected, this evidence is not conclusive.

A significant result is only a piece of evidence, and there should always be additional criteria in decision making and research.

4. Proper inference requires full reporting and transparency.

Before conducting research and before collecting data, the experiment needs to be designed and hypotheses need to be stated. Often, especially with a dramatic increase in access to “Big Data”, some have used data dredging as a way to look at many possibilities and identify phenomena that are significant. Researchers, in a desire to get published, will cheat the science by using techniques called **p -hacking**.

Methods of p -hacking

- Collecting data until the p -value $< \alpha$, then stop collecting data.
- Analyzing many options or conditions, but only publishing ones that are significant.
- Cherry picking the data to only include values that support the claim.
- Only looking at subgroups that are significant.

Use of these p -hacking methods are troubling and is one of the main reasons scientific journals are now skeptical of p -value based hypothesis testing.

The XKCD comic “Significant”⁷⁹, pictured on the right, shows an example of p -hacking, including how the media misinterprets research.



5. A p -value, or statistical significance, does not measure the size of an effect or the importance of a result.

A result may be statistically significant, but have no practical value.

Suppose someone claims that the mean flying time between New York and San Francisco is 6 hours 20 minutes. After conducting a large sample size study, you find significant evidence (p -value < .01) that the mean flying time is really longer, with a sample mean

of 6 hours and 23 minutes.

Even though your evidence is strong, there is no practical difference between the times. The p -value does not address effect sizes.

6. By itself, a p -value does not provide a good measure of evidence regarding a model or hypothesis.

The p -value is a useful tool, but by itself is not enough to support research.⁸⁰

<u>P-VALUE</u>	<u>INTERPRETATION</u>
0.001	HIGHLY SIGNIFICANT
0.01	
0.02	
0.03	
0.04	SIGNIFICANT
0.049	
0.050	OH CRAP REDO CALCULATIONS.
0.051	ON THE EDGE OF SIGNIFICANCE
0.06	
0.07	HIGHLY SUGGESTIVE, SIGNIFICANT AT THE $P < 0.10$ LEVEL
0.08	
0.09	
0.099	HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS
≥ 0.1	

This page titled [14.9: The \$p\$ -value- Misconceptions and Proper Usage](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Maurice A. Geraghty](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [10.9: The \$p\$ -value- Misconceptions and Proper Usage](#) by [Maurice A. Geraghty](#) is licensed [CC BY-SA 4.0](#). Original source: <http://nebula2.deanza.edu/~mo/holisticInference.html>.

CHAPTER OVERVIEW

Chapter 15: Week 15

15.1: Review 3

15.2: Final Exam

Chapter 15: Week 15 is shared under a [CC BY](#) license and was authored, remixed, and/or curated by LibreTexts.

15.1: Review 3

15.1: Review 3 is shared under a [CC BY](#) license and was authored, remixed, and/or curated by LibreTexts.

15.2: Final Exam

15.2: Final Exam is shared under a [CC BY](#) license and was authored, remixed, and/or curated by LibreTexts.

Detailed Licensing

Overview

Title: Introduction to Probability and Mathematical Statistics

Webpages: 71

All licenses found:

- [CC BY 4.0](#): 98.6% (70 pages)
- [Undeclared](#): 1.4% (1 page)

By Page

- [Introduction to Probability and Mathematical Statistics - CC BY 4.0](#)
 - [Front Matter - CC BY 4.0](#)
 - [TitlePage - CC BY 4.0](#)
 - [InfoPage - CC BY 4.0](#)
 - [Table of Contents - Undeclared](#)
 - [Licensing - CC BY 4.0](#)
 - [Chapter 1: Week 1 - CC BY 4.0](#)
 - [1.1: Basic Set Theory - CC BY 4.0](#)
 - [1.2: The Three Axioms of Probability - CC BY 4.0](#)
 - [1.3: Basic Propositions in Probability - CC BY 4.0](#)
 - [Chapter 2: Week 2 - CC BY 4.0](#)
 - [2.1: Applications - CC BY 4.0](#)
 - [2.2: Introduction to Counting - CC BY 4.0](#)
 - [2.3: Permutations - CC BY 4.0](#)
 - [2.3.1: The Multinomial Part 1 - CC BY 4.0](#)
 - [Chapter 3: Week 3 - CC BY 4.0](#)
 - [3.1: Combinations - CC BY 4.0](#)
 - [3.2: The Multinomial - CC BY 4.0](#)
 - [3.2.1: The Multinomial Part 2 - CC BY 4.0](#)
 - [Chapter 4: Week 4 - CC BY 4.0](#)
 - [4.1: Simple Sample Spaces - CC BY 4.0](#)
 - [4.2: Overcounting - CC BY 4.0](#)
 - [Chapter 5: Week 5 - CC BY 4.0](#)
 - [5.1: Review 1 - CC BY 4.0](#)
 - [5.2: Exam 1 - CC BY 4.0](#)
 - [Chapter 6: Week 6 - CC BY 4.0](#)
 - [6.1: Conditional Probabilities - CC BY 4.0](#)
 - [6.2: Independence - CC BY 4.0](#)
 - [Chapter 7: Week 7 - CC BY 4.0](#)
 - [7.1: Bayes' Theorem - CC BY 4.0](#)
 - [Chapter 8: Week 8 - CC BY 4.0](#)
 - [8.1: Discrete Random Variables - CC BY 4.0](#)
 - [8.2: The Cumulative Distribution Function - CC BY 4.0](#)
 - [8.3: Expected Values - CC BY 4.0](#)
 - [Chapter 9: Week 9 - CC BY 4.0](#)
 - [9.1: Variances - CC BY 4.0](#)
 - [9.2: The Uniform Random Variable - CC BY 4.0](#)
 - [9.3: The Bernoulli and Binomial Random Variables - CC BY 4.0](#)
 - [9.4: The Geometric and Negative Binomial Random Variables - CC BY 4.0](#)
 - [9.5: The Poisson Random Variable - CC BY 4.0](#)
 - [Chapter 10: Week 10 - CC BY 4.0](#)
 - [10.1: Review 2 - CC BY 4.0](#)
 - [10.2: Exam 2 - CC BY 4.0](#)
 - [Chapter 11: Week 11 - CC BY 4.0](#)
 - [11.1: Continuous Random Variables - CC BY 4.0](#)
 - [11.2: The Uniform Continuous Distribution - CC BY 4.0](#)
 - [11.3: The Exponential Distribution - CC BY 4.0](#)
 - [11.4: The Normal Distribution - CC BY 4.0](#)
 - [Chapter 12: Week 12 - CC BY 4.0](#)
 - [12.1: The Central Limit Theorem - CC BY 4.0](#)
 - [Chapter 13: Week 13 - CC BY 4.0](#)
 - [13.1: Inferential Statistics - CC BY 4.0](#)
 - [13.2: Confidence Intervals - CC BY 4.0](#)
 - [Chapter 14: Week 14 - CC BY 4.0](#)
 - [14.1: Procedures of Hypotheses Testing and the Scientific Method - CC BY 4.0](#)
 - [14.2: Formulate General Research Questions - CC BY 4.0](#)
 - [14.3: Design Research Hypotheses and Experiment - CC BY 4.0](#)
 - [14.4: Collect and Analyze Experimental Data - CC BY 4.0](#)
 - [14.5: Report Conclusions in Non-statistical Language - CC BY 4.0](#)
 - [14.6: Test of Mean vs. Hypothesized Value – A Complete Example - CC BY 4.0](#)
 - [14.7: Type II Error and Statistical Power - CC BY 4.0](#)
 - [14.8: New Models for One Population Inference, Similar Procedures - CC BY 4.0](#)

- 14.9: The p-value- Misconceptions and Proper Usage
- [CC BY 4.0](#)
- Chapter 15: Week 15 - [CC BY 4.0](#)
 - 15.1: Review 3 - [CC BY 4.0](#)
 - 15.2: Final Exam - [CC BY 4.0](#)
- Back Matter - [CC BY 4.0](#)
 - Index - [CC BY 4.0](#)
 - Glossary - [CC BY 4.0](#)
 - Detailed Licensing - [CC BY 4.0](#)
 - Detailed Licensing - [CC BY 4.0](#)

Detailed Licensing

Overview

Title: Introduction to Probability and Mathematical Statistics

Webpages: 71

All licenses found:

- [CC BY 4.0](#): 98.6% (70 pages)
- [Undeclared](#): 1.4% (1 page)

By Page

- [Introduction to Probability and Mathematical Statistics - CC BY 4.0](#)
 - [Front Matter - CC BY 4.0](#)
 - [TitlePage - CC BY 4.0](#)
 - [InfoPage - CC BY 4.0](#)
 - [Table of Contents - Undeclared](#)
 - [Licensing - CC BY 4.0](#)
 - [Chapter 1: Week 1 - CC BY 4.0](#)
 - [1.1: Basic Set Theory - CC BY 4.0](#)
 - [1.2: The Three Axioms of Probability - CC BY 4.0](#)
 - [1.3: Basic Propositions in Probability - CC BY 4.0](#)
 - [Chapter 2: Week 2 - CC BY 4.0](#)
 - [2.1: Applications - CC BY 4.0](#)
 - [2.2: Introduction to Counting - CC BY 4.0](#)
 - [2.3: Permutations - CC BY 4.0](#)
 - [2.3.1: The Multinomial Part 1 - CC BY 4.0](#)
 - [Chapter 3: Week 3 - CC BY 4.0](#)
 - [3.1: Combinations - CC BY 4.0](#)
 - [3.2: The Multinomial - CC BY 4.0](#)
 - [3.2.1: The Multinomial Part 2 - CC BY 4.0](#)
 - [Chapter 4: Week 4 - CC BY 4.0](#)
 - [4.1: Simple Sample Spaces - CC BY 4.0](#)
 - [4.2: Overcounting - CC BY 4.0](#)
 - [Chapter 5: Week 5 - CC BY 4.0](#)
 - [5.1: Review 1 - CC BY 4.0](#)
 - [5.2: Exam 1 - CC BY 4.0](#)
 - [Chapter 6: Week 6 - CC BY 4.0](#)
 - [6.1: Conditional Probabilities - CC BY 4.0](#)
 - [6.2: Independence - CC BY 4.0](#)
 - [Chapter 7: Week 7 - CC BY 4.0](#)
 - [7.1: Bayes' Theorem - CC BY 4.0](#)
 - [Chapter 8: Week 8 - CC BY 4.0](#)
 - [8.1: Discrete Random Variables - CC BY 4.0](#)
 - [8.2: The Cumulative Distribution Function - CC BY 4.0](#)
 - [8.3: Expected Values - CC BY 4.0](#)
 - [Chapter 9: Week 9 - CC BY 4.0](#)
 - [9.1: Variances - CC BY 4.0](#)
 - [9.2: The Uniform Random Variable - CC BY 4.0](#)
 - [9.3: The Bernoulli and Binomial Random Variables - CC BY 4.0](#)
 - [9.4: The Geometric and Negative Binomial Random Variables - CC BY 4.0](#)
 - [9.5: The Poisson Random Variable - CC BY 4.0](#)
 - [Chapter 10: Week 10 - CC BY 4.0](#)
 - [10.1: Review 2 - CC BY 4.0](#)
 - [10.2: Exam 2 - CC BY 4.0](#)
 - [Chapter 11: Week 11 - CC BY 4.0](#)
 - [11.1: Continuous Random Variables - CC BY 4.0](#)
 - [11.2: The Uniform Continuous Distribution - CC BY 4.0](#)
 - [11.3: The Exponential Distribution - CC BY 4.0](#)
 - [11.4: The Normal Distribution - CC BY 4.0](#)
 - [Chapter 12: Week 12 - CC BY 4.0](#)
 - [12.1: The Central Limit Theorem - CC BY 4.0](#)
 - [Chapter 13: Week 13 - CC BY 4.0](#)
 - [13.1: Inferential Statistics - CC BY 4.0](#)
 - [13.2: Confidence Intervals - CC BY 4.0](#)
 - [Chapter 14: Week 14 - CC BY 4.0](#)
 - [14.1: Procedures of Hypotheses Testing and the Scientific Method - CC BY 4.0](#)
 - [14.2: Formulate General Research Questions - CC BY 4.0](#)
 - [14.3: Design Research Hypotheses and Experiment - CC BY 4.0](#)
 - [14.4: Collect and Analyze Experimental Data - CC BY 4.0](#)
 - [14.5: Report Conclusions in Non-statistical Language - CC BY 4.0](#)
 - [14.6: Test of Mean vs. Hypothesized Value – A Complete Example - CC BY 4.0](#)
 - [14.7: Type II Error and Statistical Power - CC BY 4.0](#)
 - [14.8: New Models for One Population Inference, Similar Procedures - CC BY 4.0](#)

- 14.9: The p-value- Misconceptions and Proper Usage
- *CC BY 4.0*
- Chapter 15: Week 15 - *CC BY 4.0*
 - 15.1: Review 3 - *CC BY 4.0*
 - 15.2: Final Exam - *CC BY 4.0*
- Back Matter - *CC BY 4.0*
 - Index - *CC BY 4.0*
 - Glossary - *CC BY 4.0*
 - Detailed Licensing - *CC BY 4.0*
 - Detailed Licensing - *CC BY 4.0*