Introduction to Statistical Inquiry

Adel Mohammadpour
University of Calgary
adel.mohammadpour@ucalgary.ca

SeyedParsa Hosseinipour Rafsanjani
parsahosseini2001@gmail.com

May 2, 2025

This book is published under a Creative Commons BY-SA license (CC BY-SA) version 4.0. This means that this book can be reused, remixed, retained, revised and redistributed (including commercially) as long as appropriate credit is given to the authors. If you remix, or modify the original version of this open textbook, you must redistribute all versions of this open textbook under the same license - CC BY-SA.

https://creativecommons.org/licenses/by-sa/4.0/

Table of Contents

1 Probability and Risk **1**
1.1 Indtroduction 1
1.2 Random Experiment and Sample Space 2
1.3 Event 3
1.4 Set Theory Operations on Events 4
1.5 What is Probability? 8
1.6 Uniform Probability Model 13
1.7 Conditional Probability 15
1.8 Contingency Tables 18
1.9 Independent Events 20
1.10 Bayes' Theorem 21
1.11 Sensitivity, Specificity, Prevalence, and Relative Risk 24

2 Probability and Accuracy **29**
2.1 Probability 29
2.2 Conditional Probability & Bayes' Theorem 29
2.3 Prevalence, Relative Risk, Sensitivity, and Specificity 29

3 Descriptive Statistics **31**
3.1 Introduction to Data 31
3.2 Exploratory Data Analysis 31

4 Probability Distributions **33**
4.1 Random Variables 33
4.2 Binomial Distribution 33
4.3 Normal Distribution 33
4.4 Sampling Distribution and Central Limit Theorem 33

5 Estimation **35**
5.1 Point Estimation 35
5.2 Confidence Interval 35

6 Estimation **37**
6.1 Hypothesis Testing for One Sample 37
6.2 Hypothesis Testing for Two Variances 37
6.3 Hypothesis for Two Means 37
6.4 Hypothesis for Two Proportions 37

7 Hypothesis Testing **39**

8 **41**

9 Probability and Accuracy **43**

10 **45**

11 **47**

12 **49**

13 **51**

14 **53**

15 **55**

16 **57**

17 Probability and Accuracy **59**
17.1 Probability 59
17.2 Conditional Probability & Bayes' Theorem 59
17.3 Prevalence, Relative Risk, Sensitivity, and Specificity 59

18 Descriptive Statistics **61**
18.1 Introduction to Data 61
18.2 Exploratory Data Analysis 61

19 Probability Distributions **63**
19.1 Random Variables 63
19.2 Binomial Distribution 63
19.3 Normal Distribution 63
19.4 Sampling Distribution and Central Limit Theorem 63

1. Probability and Risk

> *"When there are but two players, your theory which proceeds by combinations is very just. But when there are three, I believe I have a proof that it is unjust that you should proceed in any other manner than the one I have."*
>
> – Pascal's letter to Fermat[1]

1.1

Indtroduction

While scientists have always tried to understand the universe with technology and explain it with complete certainty, this is not always possible. In other words, there is no other way but to accept chance as part of our lives.

The concept of chance has been extensively explored across elementary, professional, and philosophical literature, serving as a key motivation for our study. While chance often appears unpredictable and devoid of structure, mathematicians have long sought to define it through rules and systematic frameworks. Not surprisingly, gamblers were among the first to seek systematic frameworks for understanding their games - probing the mechanics of luck, wins and losses. In 1654, a Parisian gambler named Antoine Gombaud (alias Chevalier de Méré), posed critical questions about winning probabilities to two of the era's greatest mathematicians: Blaise Pascal and Pierre de Fermat. Through their correspondence, the two mathematicians initiated the development of modern probability theory. However, Gerolamo Cardano and Galileo Galilei, two Italian scholars, had also made significant contributions that captured the interest of Italian gamblers.

After years of development, the Russian mathematician Andrey Kolmogorov introduced the standard probability axioms in 1933, establishing the rigorous foundations of modern probability theory. Today, probability theory serves as a fundamental tool across diverse fields including social sciences, medicine, biology, machine learning, physics, and countless other applications. The following sections provide rigorous definitions of key concepts needed to establish a precise theoretical foundation.

---

[1] from `https://www.york.ac.uk/depts/maths/histstat/pascal.pdf`

1.2

Random Experiment and Sample Space

Suppose we want to conduct an experiment for which we know all possible outcomes. Assuming the experimental conditions remain constant each time, if each realization of this experiment produces exactly one outcome, we call it a **random experiment**. We denote the set of all possible outcomes of such an experiment by S, called the **sample space**, which is a nonempty set.

Example 1.2.1 *Consider the experiment of tossing a coin where the experimental conditions are controlled such that the coin lands on heads or tails, with no other possible outcomes. For instance the surface is chosen so that it won't land on edge. This is a random experiment in which we are interested in observing whether the coin lands heads or tails when viewed from above after it comes to rest. Hence, denoting the outcome of observing heads by* H *and tails by* T*, the sample space of this experiment is* $S = \{H, T\}$.

Example 1.2.2 *Consider an experiment where a ball is drawn from an urn containing one blue, one green, and one red ball. The experimenter cannot see inside the urn when making each draw. This is a random experiment in which we are interested in the color of the drawn ball. Hence, denoting the outcome of observing blue, green, and red ball by* B, G*, and* R *respectively, the sample space of this experiment is* $S = \{B, G, R\}$.

Example 1.2.3 *Consider the experiment of drawing a card from a deck. Each time a card is drawn, it is returned to the deck and the duck is shuffled, so the conditions of each experiment remains the same. If we are interested in the suit of the drawn cards, the sample space is* $S = \{\clubsuit, \blacklozenge, \spadesuit, \heartsuit\}$. *But if we are interested in the rank of the cards, the sample space becomes* $S = \{A, 1, 2, 3, 4, 5, 6, 7, 8, 9, J, Q, K\}$.

In some cases, as the following example demonstrates, the sample space may be infinite:

Example 1.2.4 *A coin is repeatedly tossed under the experimental conditions of Example 1.2.1 until heads appears. There are infinitely many possible outcomes: If the first coin toss results in heads, the experiment is immediately terminated. If it's tails, the coin is tossed again. The experiment terminates if heads appears for the second time. If not, the experiment continues until heads is observed. So the sample space in this case is* $S = \{H, TH, TTH, TTTH, \dots\}$.

The first outcome is the case where heads appears on the first coin toss. The second outcome corresponds to heads appearing on the second coin toss, and so on.

The sample space can also be continuous, as demonstrated in the following example:

Example 1.2.5 *Consider an experiment to determine the lifespan (in seconds) of lightbulbs produced under identical manufacturing conditions. The sample space* S *consists of all possible positive real numbers representing time until lightbulb failure, i.e.,* $S = \{t \in R | t \geq 0\}$.

1.3 Event

Each outcome of a random experiment is an element of its sample space, S. In doing these experiments, we are interested in observing some specific outcomes, or in other words, a subset of S. This subset is called an **event** and is denoted by E. But we know that in each realization of a random experiment, only one element $e \in S$ is observed. If $e \in S$, we say E has occurred and if $e \in S \setminus E$, we say it has not occurred.

Example 1.3.1 *In the experiment of throwing a six-sided die,* $S = \{1, 2, 3, 4, 5, 6\}$. *In case the experimenter is interested in the event of observing an even number, the event of interest would be* $E = \{2, 4, 6\}$*, a subset of* S*. If instead they are interested in the event of observing an odd number, the event of interest would become* $E = \{1, 3, 5\}$*, again a subset of* S.

Note that in the previous example, we omitted certain details about the experimental conditions and our observations of interest. Henceforth, unless explicitly stated otherwise, we make the following standard assumptions:

- Each realization of a random experiment is performed under identical conditions.
- For a die roll, the random experiment of interest is observing the uppermost face after landing on the ground.
- For a coin toss, the random experiment of interest is whether it lands heads (H) or tails (T).

If $E = S$, the event is called a **sure event**, since each element of S is in E and so observing any outcome means the event has occurred. On the other hand, if $E = \emptyset$, the event is called an **impossible event**, since it contains no element at all.

Example 1.3.2 *In Example 1.3.1, observing a number greater than 6 is an impossible event, while observing a number less than 7 is a sure event.*

An event consisting of only a single element of the sample space S is called an **elementary event** or an **atomic event**.

Example 1.3.3 *In Example 1.3.1,* $\{1\}, \{2\}, \{3\}, \{4\}, \{5\}$*, and* $\{6\}$ *are all elementary events.*

For two events E and F, if every element of E is also in F, we say that E is a **subevent** of F, or in set notation, $E \subset F$.

Example 1.3.4 *In Example 1.3.1, if E is the event "observing a number greater than 5" and F is the event "observing a number greater than 4", then $E = \{6\}$ and $F = \{5, 6\}$, and thus $E \subset F$. It is trivial that the occurrence of E implies the occurrence of F, but not conversely. For instance, if the die roll outcome is 5, F occurs but E does not.*

Two events are called **equal events** if they consist of same elements. By definition, if $E \subset F$ and $F \subset E$, then E and F are equal events and we write $E = F$.

Example 1.3.5 *In Example 1.2.3, suppose the sample space is $S = \{A, 1, 2, 3, 4, 5, 6, 7, 8, 9, J, Q, K\}$. If E is the event "observing a number less than 5 in Clubs" and F is the event "observing a number less than 5 in Diamonds", then $E = \{1, 2, 3, 4\}$ and $F = \{1, 2, 3, 4\}$, and thus $E = F$. Trivially, the occurrence of E implies the occurrence of F, and vice versa.*

1.4

Set Theory Operations on Events

We denoted the set of all outcomes of a random experiment by S, and we saw that we may be interested in a subset of this set, $E \subset S$. Now consider two events $E, F \subset S$. From these two events, we can derive additional events as demonstrated in this section.

1.4.1 Union of Two Events

The **union** of two events E and F consists of all elements that are in E or in F and is denoted by $E \cup F$. The occurrence of E or F results in the occurrence of $E \cup F$.

Example 1.4.1 *In Example 1.3.1, we denote the event of observing an even number by E and the event of observing a prime number by F. So $E = \{2, 4, 6\}$ and $F = \{2, 3, 5\}$. The union of these two events consists of all elements in E or F, meaning $E \cup F = \{2, 3, 4, 5, 6\}$. So if the outcome of a die roll is 4, we say the event "observing a number which is even or prime" occurs, since 4 is an even number.*

If it's 5, we also say that this event occurs, since 5 is a prime number.

Now suppose we observe 2 in a die roll. We again say that $E \cup F$ occurs. The important thing to note here is that this "or" is an inclusive or, which means $E \cup F$ is actually "observing a number that is even or prime or both". In this case, 2 is both an even and a prime number.

1.4.2 Intersection of Two Events

The **intersection** of two events E and F consists of all elements in both E and F and is denoted by $E \cap F$. In order for $E \cap F$ to occur, both E and F has to occur.

Example 1.4.2 *Consider the events E and F in Example 1.4.1. $E \cap F$ is the event "observing a number that is both even and prime". In other words, $E \cap F$ consists of all elements that are in both E and F, and so are both even and prime. Hence, $E \cap F = \{2\}$.*

So if 3 is observed in rolling a die, $E \cap F$ does not occur since 3 is not even while it is a prime number. If 4 is observed, $E \cap F$ again does not occur, since 4 is even but not prime. The only acceptable observation to say $E \cap F$ occurs is 2, since only 2 is both even and prime among all the elements of E and F.

If $E \cap F = \varnothing$, we say the two events are **disjoint** or **mutually exclusive**. For instance, in Example 1.4.1, if we denote the event "observing an odd number" by O, $E \cap O = \varnothing$, meaning E and O are mutually exclusive. In other words, a single die roll cannot simultaneously result in both an even and an odd number.

Note that events $A_1, A_2, \ldots$ are said to be disjoint (or mutually exclusive) if $A_i \cap A_j = \varnothing$ for $i \neq j$.

1.4.3 Difference of Two Events

The **difference** $E \setminus F$ is the event containing all elements in E but not in F.

Example 1.4.3 *In Example 1.4.1, $E \setminus F$ is the event "observing a number that is even but not prime". Thus $E \setminus F = \{4, 6\}$. $3 \notin E - \setminus F$ since it is not even, and $2 \notin E - F$ because while it is an even number, it is also a prime number. So observing 3 or 2 upon rolling a die means $E \setminus F$ does not occur. For $E \setminus F$ to occur, we have to observe either 4 or 6.*

1.4.4 Complement of an Event

The **complement** of an event E is the difference $S \setminus E$, containing all alements in sample space that are not in E. The complement of E is denoted by E^{C}, E', or $\bar{E}$.

Example 1.4.4 *In Example 1.4.1, E^{C} is the event "observing a number that is not even", which means $E^{\mathsf{C}} = \{1, 3, 5\}$. So E^{C} occurs only when the observed number on the die is odd.*

Note that "observing a number that is not even" does not mean any number that is not even, but rather any number in the sample space $S = \{1, 2, 3, 4, 5, 6\}$ that is not even.

1.4.5 Symmetric Difference of Two Events

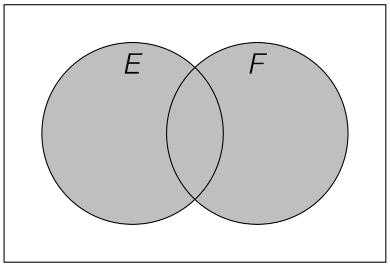
The **symmetric difference** of two events E and F consists of all elements in E or F but not in both and is denoted by $E \bigtriangleup F$. By definition, $E \bigtriangleup F = F \bigtriangleup E$ and this is why this operation is called symmetric.

Example 1.4.5 *In Example 1.4.1,* $E \bigtriangleup F = \{3, 4, 5, 6\}$*, containing numbers in the sample space that are even or prime but not both. Since 2 is both even and prime, it follows that* $2 \in E \cup F$*, but* $2 \notin E \bigtriangleup F$*.*

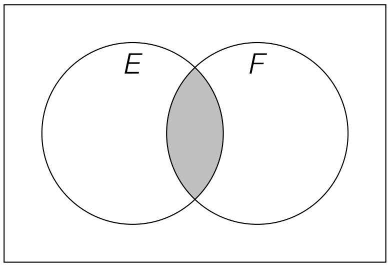
Note the exclusive "or" in this operation, unlike the inclusive case discussed in Example 1.4.1.

The set operations and key concepts for two events E and F are visually represented on this page. In these diagrams, the sample space is depicted as a rectangle, events E and F as two circles, and the operation or concept specified above each rectangle is highlighted through coloring.

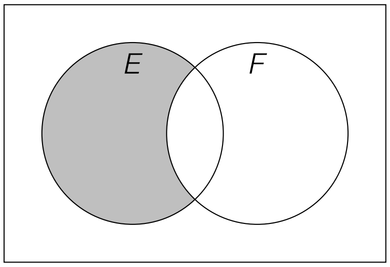
$E \cup F$



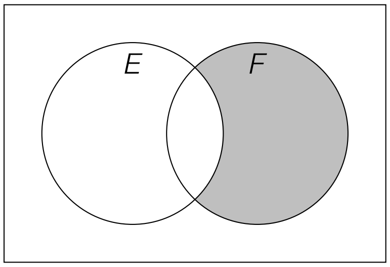
$E \cap F$



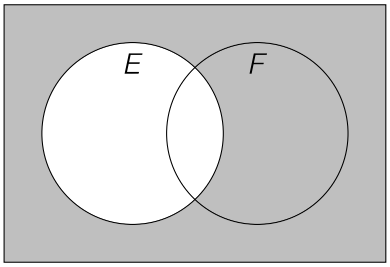
$E \setminus F$



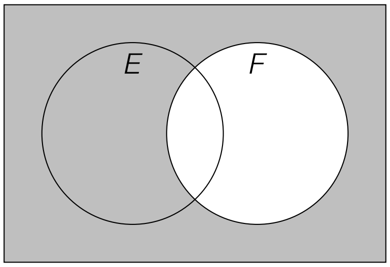
$F \setminus E$



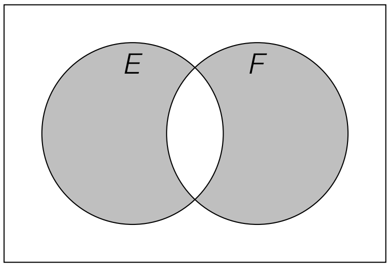
E^{C}



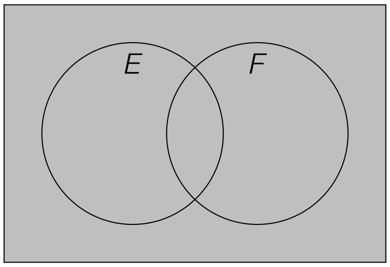
F^{C}



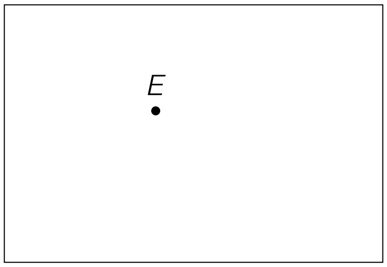
$E \triangle F$



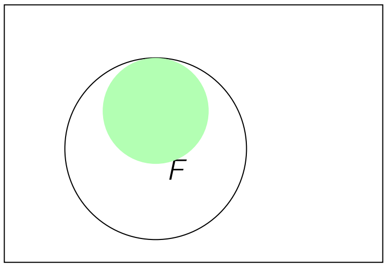
S



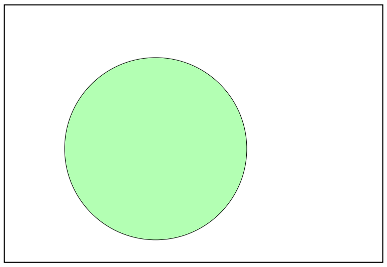
$E = \varnothing$



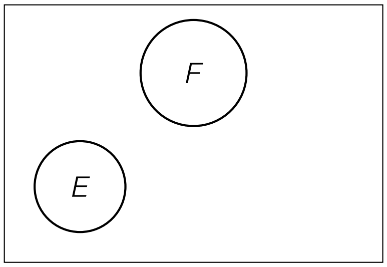
$F \subseteq E$



$E = F$



Disjoint



1.5

What is Probability?

There are three prominent ways to define probability. In the three following subsections, these three definitions are explored. Then an axiomatic definition of probability is given which is consistent with these three definitions.

1.5.1 Classical Definition of Probability

In this definition, each elementary event of the sample space is considered to be equally likely to occur. For example in rolling a fair die, all the events $\{1\}, \{2\}, \{3\}, \{4\}, \{5\}$ and $\{6\}$ have the same probability. Or in tossing a coin (the random experiment in Example 1.2.1), the elementary events $\{H\}$ and $\{T\}$ are both equally likely to happen, which is consistent with our intuition of 50% chance of occurring for each event.

This interpretation of probability goes back to Pierre-Simon Laplace, who wrote:

> The probability of an event is the ratio of the number of cases favorable to it, to the number of all cases possible when nothing leads us to expect that any one of these cases should occur more than any other, which renders them, for us, equally possible. [2]

So in this interpretation of probability, for an event $E \subset S$, the probability of E is given by:

$$P(E) = \frac{\text{number of outcomes in } E}{\text{total number of outcomes in } S}$$

While we can assign a probability to each elementary event when the sample space consists of finitely many elements, it is not possible when the elementary events are not equally likely to occur or the sample space is infinite like we saw in Example 1.2.4.

1.5.2 Frequentist Definition of Probability

Consider a randon experiment in which we are interested in the occurrence of event E. Suppose this experiment is repeated n times under identical experimental conditions, and event E occurs r times in total. r and $\frac{r}{n}$ are said to be the **frequency** and **relative frequency** of E in these n trials, respectively.

As the number of trials n increases, the frequency r and consequently the relative frequency $\frac{r}{n}$ changes as well. However, empirical observations show that the relative frequency converges to a constant value, which in this interpretation is defined as the probability of E.

---

[2] Laplace, Théorie analytique des probabilités, retrieved from `https://en.wikipedia.org/wiki/Classical_definition_of_probability`

The frequentist interpretation of probability may trace its earliest conceptual origins to Aristotle, who wrote:

> the probable is that which for the most part happens [3]

The frequentist interpretation of probability rose to dominance during the 19th century, becoming the foundation of classic statistical inference.

Let us reconsider the random experiment described in Example 1.2.2. According to the classical interpretation of probability, the probability of drawing a blue ball is $\frac{1}{3}$. However, if we don't know the urn's contents, we can determine the probability of drawing a blue ball through a simple experiment. By repeatedly drawing balls from the urn and calculating the relative frequency of drawing blue balls, we observe that this ratio converges to $\frac{1}{3}$ as the number of trials increases, which is consistent with the classical definition of probabiltiy. But the frequentist interpretation offers another advantage: suppose there are two blue balls, one green ball, and one red ball in the urn. Here, while the sample space remains $\{B, G, R\}$, the elementary events now have different probabilities because there is an additional blue ball in the urn! So while we cannot calculate the probability of drawing a blue ball using the classical definition, we can determine it through the frequentist approach by performing many trials and observing the convergence of relative frequency.

While this interpretation sounds like a better approach to defining probability, it faces several limitations. Consider, for instance, estimating the probability of precipitation occurring tomorrow. There is, in fact, only one tomorrow; we cannot conduct multiple trials of "tomorrow" to count rainy occurrences and determine their relative frequency. Moreover, maintaining truly identical experimental conditions is practically impossible in most real-world scenarios. Another fundamental challenge lies in determining how many trials are sufficient for the relative frequency to converge to a stable probability value.

1.5.3 Epistemic Definition of Probability

In this interpretation of probability, each observer assigns a subjective probability to an event based on their prior beliefs. Reconsider the urn example from the previous subsection. The experimenter draws a ball from the urn without revealing it to the observers.

An observer who saw the experimenter add an extra blue ball to the urn assigns a probability of $\frac{2}{4}$ to the ball being blue. Another observer, who previously knew the urn contained one blue, one green, and one red ball, assigns a probability of $\frac{1}{3}$. The experimenter, however, knows with certainty whether the ball is blue or not.

In the frequentist approach, the ball is either blue or not, and the probability is determined through repeated trials. Thus, frequentists don't assign probabilities to single drawn balls. In the epistemic perspective, however, each individual assigns a probability based on their existing knowledge.

---

[3] Aristotle, Rhetoric, retrieved from `https://en.wikipedia.org/wiki/Frequentist_probability`

1.5.4 Axiomatic Definition of Probability

Just like Euclidean geometry, where theorems are derived from a set of axioms taken as true[4], modern probability theory is built upon axioms proposed by the Russian mathematician Andrey Kolmogorov in 1933.

A **probability measure** on a sample space S is a function $P(.)$ that assigns to each event $E \subset S$ a real number $P(A)$, called the probability of A, satisfying the following three **axioms of probability**:

- for each event $E \subset S$, the probability of A is non-negative, meaning $P(E) \geq 0$.
- the probability of S is 1, meaning $P(S) = 1$.
- if $E_1, E_2, \ldots$ is a sequence of disjoint events, meaning $E_i \cap E_j = \varnothing$ for each $i \neq j$, then $P(E_1 \cup E_2 \cup \ldots) = P(E_1) + P(E_2) + \ldots$.

The first axiom states that probabilities cannot be negative, which aligns with our intuition. The second axiom indicates that since every experimental outcome belongs to the sample space, the event S must occur with probability 1. The third axiom establishes that for any countably infinite number of disjoint events, the probability of their union equals the sum of their individual probabilities.

Different interpretations of probability remain consistent with the axiomatic definition. Consider the frequentist approach:

- Relative frequencies are non-negative, satisfying the first axiom.
- The relative frequency of S equals 1 in any number of trials, since every outcome belongs to the sample space.
- For mutually exclusive events, the relative frequency of their union equals the sum of their individual relative frequencies, as each outcome is counted only once.

Next, we prove some thorems using the axioms of probability.

Theorem 1.5.1 *The probabiliy of impossible event, $\varnothing$, is zero.*

Proof: Consider the sequence of events $E_1, E_2, \ldots$ where $E_1 = S$ and $E_i = \varnothing$ for $i > 1$. Since the events in this sequence are mutually exclusive and $S = \bigcup_{i=1}^{\infty} E_i$, by the third axiom we have

$$P(S) = P(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i) = P(S) + \sum_{i=2}^{\infty} P(\varnothing) = P(S) + P(\bigcup_{i=2}^{\infty} \varnothing) = P(S) + P(\varnothing)$$

Cancelling $P(S)$ from both sides we obtain $P(\varnothing) = 0$. □

---

[4] https://www.math.brown.edu/tbanchof/Beyond3d/chapter9/section01.html

Theorem 1.5.2 *For finitely many disjoint events $E_1, E_2, \dots, E_n$, we have*

$$P(\bigcup_{i=1}^{n} E_i) = \sum_{i=1}^{n} P(E_i)$$

Proof: Consider the sequence of events $E_1, E_2, \dots$ where $E_i = \varnothing$ for $i > n$. Since all events in this sequence are mutually exclusive and $\bigcup_{i=1}^{n} E_i = \bigcup_{i=1}^{\infty} E_i$, by applying the third axiom and Theorem 1.5.1, we obtain:

$$\begin{aligned}
P(\bigcup_{i=1}^{n} E_i) = P(\bigcup_{i=1}^{\infty} E_i) &= \sum_{i=1}^{\infty} P(E_i) \\
&= \sum_{i=1}^{n} P(E_i) + \sum_{i=n+1}^{\infty} P(E_i) \\
&= \sum_{i=1}^{n} P(E_i) + P(\bigcup_{i=n+1}^{\infty} E_i) \\
&= \sum_{i=1}^{n} P(E_i) + P(\varnothing) \\
&= \sum_{i=1}^{n} P(E_i)
\end{aligned}$$

□

Theorem 1.5.3 *For any event $E \subset S$, $P(E^{\mathsf{C}}) = 1 - P(E)$.*

Proof: E and E^{C} are mutually exclusive and $S = E \cup E^{\mathsf{C}}$ since every element of the sample space is either in E or not. Thus, using the second axiom and Theorem 1.5.2, we have

$$1 = P(S) = P(E \cup E^{\mathsf{C}}) = P(E) + P(E^{\mathsf{C}})$$
$$P(E^{\mathsf{C}}) = 1 - P(E)$$

□

Corollary 1.5.3.1 *Using De Morgan's laws in set theory, for any two events $E, F \subset S$:*

$$(E \cap F)^{\mathsf{C}} = E^{\mathsf{C}} \cup F^{\mathsf{C}}$$
$$(E \cup F)^{\mathsf{C}} = E^{\mathsf{C}} \cap F^{\mathsf{C}}$$

From these we can derive:

$$P(E^{\mathsf{C}} \cup F^{\mathsf{C}}) = P((E \cap F)^{\mathsf{C}}) = 1 - P(E \cap F)$$
$$P(E^{\mathsf{C}} \cap F^{\mathsf{C}}) = P((E \cup F)^{\mathsf{C}}) = 1 - P(E \cup F)$$

Theorem 1.5.4 *For any two events $F, E \subset S$, the equality $P(F - E) = P(F) - P(E \cap F)$ holds.*

Proof: It can be shown that $F = (F - E) \cup (E \cap F)$ where $F - E$ and $E \cap F$ are mutually exclusive. By Theorem 1.5.2:

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

□

Corollary 1.5.4.1 *If $E \subset F$, then $E \cap F = E$ and thus $P(F - E) = P(F) - P(E)$.*

Corollary 1.5.4.2 *If $E \subset F$, the first axiom and Corollary 1.5.4.1 yield $P(E) \leq P(F)$. Replacing F with S and applying the second axiom gives $P(E) \leq 1$. Combined with the first axiom, this proves that for any event E,*

$$0 \leq P(E) \leq 1$$

Theorem 1.5.5 *For any two events $E, F \subset S$, the equality $P(E \cup F) = P(E) + P(F) - P(E \cap F)$ holds.*

Proof: It can be shown that $E \cup F = E \cup (F - E)$ where E and $F - E$ are mutually exclusive. By Theorem 1.5.2 and Theorem 1.5.4:

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

□

Example 1.5.1 *Suppose E, F are two events from sample space S where $P(E) = 0.6$, $P(F - E) = 0.3$ and $P(E \cap F) = 0.2$.*

1. *What is $P(F)$?*
2. *What is $P(E \cup F)$?*
3. *What is $P(E \cap F^C)$?*
4. *What is $P(E \cap E^C)$?*
5. *What is $P(E^C \cap F^C)$?*

Solution

1. From Theorem 1.5.4, we derive:

$$P(F) = P(F - E) + P(E \cap F) = 0.3 + 0.2 = 0.5$$

2. By Theorem 1.5.5:

$$P(E \cup F) = P(E) + P(F) - P(E \cap F) = 0.6 + 0.5 - 0.2 = 0.9$$

3. From set theory, we have the equality $E \cap F^C = E - F$. Intuitively, $E \cap F^C$ and $E - F$ both contain exactly those elements of S that belong to E but not to F. Applying Theorem 1.5.4 yields:

$$P(E \cap F^C) = P(E - F) = P(E) - P(F \cap E) = P(E) - P(E \cap F) = 0.6 - 0.2 = 0.4$$

4. No element of S can simultaneously belong to E and not belong to it , thus $E \cap E^C = \varnothing$. Theorem 1.5.1 implies:

$$P(E \cap E^C) = P(\varnothing) = 0$$

5. By Corollary 1.5.3.1:

$$P(E^C \cap F^C) = 1 - P(E \cup F) = 1 - 0.9 = 0.1$$

■

Exercise 1.1 *Show that for any two events* $E, F \subset S$, $P(E^C \cup F) = 1 - P(E) + P(F \cap E)$.

Exercise 1.2 *Suppose* E, F *are two events of sample space* S.

1. *Show that* $P(E \cap F) + P(E \cap F^C) = P(E)$.
2. *Show that* $P(E) + P(E^C) = 1$.

1.6

Uniform Probability Model

In a **uniform probability model**, each elementary event has an equal probability. So if the sample space has $n(S)$ elements, then for each outcome $e_i \in S$, we have:

$$P(\{e_i\}) = \frac{1}{n(S)}$$

If an event E contains $n(E)$ elements, we can express it as $E = \bigcup_j \{e_j\}$, where each e_j is an element of S belonging to E. Applying Theorem 1.5.2:

$$P(E) = P(\bigcup_j \{e_j\}) = \sum_j P(\{e_j\}) = \sum_j \frac{1}{n(S)} = \frac{n(E)}{n(S)}$$

Comparing this with subsection 1.5.1, we observe that it is equivalent to Laplace's classical definition of probability.

Example 1.6.1 *In the coin toss random experiment from Example 1.2.1, if the event of interest E is "observing heads", then $P(E) = \frac{1}{2}$, since $E = \{H\}$ contains exactly one outcome out of two possible equally likely outcomes. Such a coin with equal probabilities of landing on heads or tails is called a* **fair coin**.

Example 1.6.2 *In the random experiment of tossing three fair coins, the sample space is $S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$. If the event of interest E is "observing at most two heads", then $E = \{HHT, HTH, HTT, THH, THT, TTH, TTT\}$, and consequently, $P(E) = \frac{7}{8}$.*

Example 1.6.3 *In the die-throw random experiment from Example 1.3.1, if the event of interest E is "observing an odd number", then $E = \{1, 3, 5\}$ and thus $P(E) = \frac{3}{6} = \frac{1}{2}$. A die with equally likely outcomes for all faces is called a* **fair die**.

Example 1.6.4 *In the random experiment of tossing two fair six-sided dice, the sample space is $S = \{(1,1), (1,2), \ldots, (6,5), (6,6)\}$. If the event of interest E is "the sum of two dice equals 8", then $E = \{(2,6), (3,5), (4,4), (5,3), (6,2)\}$ and consequently $P(E) = \frac{5}{36}$.*

Exercise 1.3 *In Example 1.6.4, what is the probability that the difference of two dice is 3?*

In the examples discussed above, we can easily count the elements in both the sample space and the event of interest. However, consider modifying Example 1.6.2 where six fair coins are tossed instead of three, or altering Example 1.6.4 to involve seven fair six-sided dice rather than two. How many elements are in the sample space in these cases? Attempting to list all possible combinations would quickly reveal what a tedious task this becomes.

There are some fundamental counting principles which aid us in these scenarios. The **multiplication principle** states that if a first task can be performed in m_1 ways, and for each of these, a second task can be performed in m_2 ways independent from the previous task, then the sequence of two tasks has $m_1 \times m_2$ possible outcomes. This naturally extends to n sequential tasks, where the total number of possible outcomes becomes $m_1 \times m_2 \times \ldots \times m_n$.

Example 1.6.5 *A standard deck contains cards with four suits (♣, ♦, ♠, ♥) and thirteen ranks (A, 1, 2, ..., 9, J, Q, K). How many total cards are in such a deck?*

Solution We decompose the counting problem into two sequential tasks: selecting a suit and then selecting a rank. There are four ways to do the former and thirteen ways to do the latter, so there are $4 \times 13 = 52$ such cards in total. ■

The **addition principle** states that if one task can be performed in n_1 ways and another distinct task in n_2 ways, then either task can be done in $n_1 + n_2$ total ways. For example, with three blue garments and seven green ones, selecting either a blue or green garment can be done in $3 + 7 = 10$ ways. This result can be generalized to more than two tasks, analogous to the extension of the multiplication principle.

Example 1.6.6 *From a group of seven statisticians and five computer scientists, two people are randomly selected for a project. What is the probability that the team consists of one statistician and one computer scientist?*

Solution The first person is chosen from 12 individuals, and the second from the remaining 11. So the sample space has $n(S) = 12 \times 11 = 132$ elements. The event E contains elements of S with one statistician and one computer scientist. So to count the number of elements in E, we again decompose the problem into two tasks. First choosing one statistician and second choosing a computer scientist, which can be done in $7 \times 5 = 35$ ways. Or first choosing one computer scientist and then choosing a statistician, which can be done in $5 \times 7 = 35$ ways. Since we have to choose a statistician first and then a computer scientist, or first a computer scientist and then a statistician, the addition principle implies $n(E) = 35 + 35 = 70$, yielding $P(E) = \frac{70}{132} = \frac{35}{66}$. ■

1.7

Conditional Probability

In previous sections, we examined basic probability examples. But sometimes in our problems, we have some prior information. In Example 1.6.4, suppose we know beforehand that one die has come up 2. Here, the event of interest becomes $E \cap F$, where F is the event "one die comes up 2". So we are interested in the probability of both E and F occurring, which is $E \cap F = \{(2,6),(6,2)\}$. On the other hand, since we have information that one die has come up 2, the sample space is restricted from S to F. So the probability of E given we know F has occurred is $\frac{n(E \cap F)}{n(F)}$. By dividing the numerator and denominator by $n(S)$, the desired probability becomes $\frac{P(E \cap F)}{P(F)}$. Since $F = \{(1,2),(2,1),(2,2),(2,3),(2,4),(2,5),(2,6),(3,2),(4,2),(5,2),(6,2)\}$, this probability is $\frac{\frac{2}{36}}{\frac{11}{36}} = \frac{2}{11}$.

The **conditional probability** of A given B, denoted $P(A|B)$, is the probability of A occurring when B is already known to have occurred. An argument similar to the preceding argument, supported by Figure 1.1, shows that $P(A|B) \propto P(A \cap B)$, which means $P(A|B) = kP(A \cap B)$ for some constant k. To determine k, observe that when $A = B$, $P(B|B) = 1$ which implies:

$$1 = kP(B \cap B) = kP(B)$$
$$k = \frac{1}{P(B)}$$

Thus, $P(A|B) = \frac{P(A \cap B)}{P(B)}$, where $P(B) > 0$. If $P(B) = 0$, $P(A|B)$ would be undefined.

Example 1.7.1 *In Example 1.6.2, let F be the event "observing exactly two tails".*

1. *What is $P(E \cap F)$?*

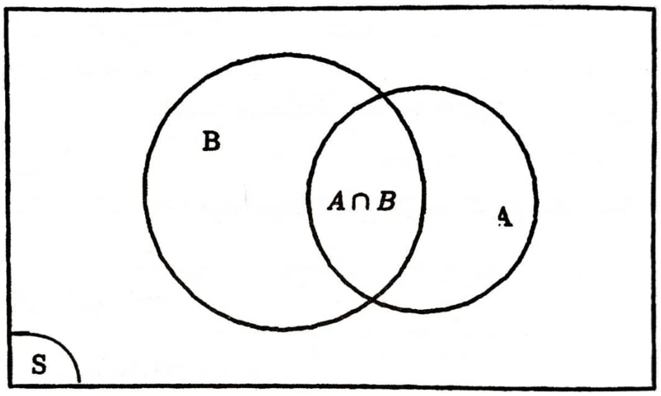


Figure 1.1: $P(A|B)$ is proportional to $P(A \cap B)$

2. *What is $P(E|F)$?*
3. *What is $P(F|E)$?*

Solution

1. First we note that $F = \{HTT, THT, TTH\}$. Thus, $E \cap F = \{HTT, THT, TTH\}$, yielding:

   $$P(E \cap F) = \frac{n(E \cap F)}{n(S)} = \frac{3}{8}$$

   Here, we are interested in the event "observing at most two heads and exactly two tails" while the sample space is not restricted to a new one since we have no prior information.

2.

   $$P(F) = \frac{n(F)}{n(S)} = \frac{3}{8}$$

   $$P(E|F) = \frac{P(E \cap F)}{P(F)} = \frac{\frac{3}{8}}{\frac{3}{8}} = 1$$

   Unlike the previous part, we now have prior information that exactly two tails are observed. By restricting our sample space to outcomes with exactly two tails, we are certain that either zero or one head will be observed, which constitutes a proper subset of the event "at most two heads", consistent with this result.

3.

   $$P(F|E) = \frac{P(F \cap E)}{P(E)} = \frac{P(E \cap F)}{P(E)} = \frac{\frac{3}{8}}{\frac{7}{8}} = \frac{3}{7}$$

Let G be the event "observing at least one tail". Since $G = E$, the two events are equal and so $P(F|E) = P(F|G)$. In other words, $P(F|E)$ is the same as the probability of observing exactly two tails given at least one tail is observed. The plausibility of observing exactly two tails after at least one tail is observed must increase, which is true since $P(F|E) = \frac{3}{7} > \frac{3}{8} = P(F)$.

■

Thus far, we have examined numerous examples of uniform probability models. However, not all probability models are uniform. Consider these cases:

Example 1.7.2 *Consider Example 1.2.2 with an urn containing three blue balls, five green balls, and two red balls. If we treat each ball as distinct, the sample space is* $S = \{B_1, B_2, B_3, G_1, G_2, G_3, G_4, G_5, R_1, R_2\}$. *This forms a uniform probability model where* $P(\{B_1\}) = P(\{B_2\}) = \ldots = P(\{R_1\}) = P(\{R_2\}) = \frac{1}{10}$. *However, if we are only interested in ball colors and not individual balls, the sample space becomes* $S = \{B, G, R\}$. *Defining* $B = \{B_1, B_2, B_3\}$, $G = \{G_1, G_2, G_3, G_4, G_5\}$ *and* $R = \{R_1, R_2\}$, *then* $P(B) = \frac{3}{10}$, $P(G) = \frac{5}{10} = \frac{1}{2}$ *and* $P(R) = \frac{2}{10} = \frac{1}{5}$, *which is a non-uniform probability model.*

Example 1.7.3 *The residents of a town have blue, green, brown, and amber eyes. When randomly selecting a resident, the event of interest is their eye color. The sample space for this experiment is* $S = \{blue, green, brown, amber\}$. *To determine the probabilities of elementary events, we conduct a survey by randomly sampling 500 people and recording their eye colors. The results are shown in the following table:*

Eye Color	**Blue**	**Green**	**Brown**	**Amber**	**Total**
Frequency	*125*	*65*	*290*	*20*	*500*
Relative Frequency	*0.25*	*0.13*	*0.58*	*0.04*	*1.00*

Note: Data reflects approximate global eye color distribution based on studies from the American Academy of Ophthalmology (2022)[5].

Using the frequentist definition of probability and the survey data above, we obtain:

$$P(\{blue\}) = 0.25, P(\{green\}) = 0.13, P(\{brown\}) = 0.58, P(\{amber\}) = 0.04$$

These unequal probabilities confirm this is not a uniform probability model.

Using this model, we can answer questions like this: If there are 362500 residents in this town, how many are expected to have green eyes?

Using the frequentist interpretation, we can answer this question through a simple calculation:

$$0.13 \times 362500 = 47125$$

The conditional probability and theorems derived from the three axioms of probability apply to any probability model. The following example demonstrates this:

---

[5] https://www.aao.org/eye-health/tips-prevention/your-blue-eyes-arent-really-blue

Example 1.7.4 *Suppose in Example 1.7.3, of people with blue eyes, 60 were female. One person is selected at random from the town. Given that the selected person has blue eyes, what is the probability that they are female?*

Solution We define event E as "the selected pesron has blue eyes" and event F as "the selected person is female". Thus:

$$P(F \cap E) = \frac{60}{500} = 0.12$$
$$P(F|E) = \frac{P(F \cap E)}{P(E)} = \frac{0.12}{0.25} = 0.48$$

■

1.8

Contingency Tables

When solving conditional probability problems, visual tools like tree diagrams and contingency tables often give invaluable insights. These techniques provide different perspectives that can clarify complex probability relationships.

Contingency tables are particularly useful when analyzing two events, say $E, F \subset S$. We can construct contingency tables using the size of events as follows:

	E	E^{C}	
F	$n(E \cap F)$	$n(E^{C} \cap F)$	$n(F)$
F^{C}	$n(E \cap F^{C})$	$n(E^{C} \cap F^{C})$	$n(F^{C})$
	$n(E)$	$n(E^{C})$	$n(S)$

Note that the sizes in margins of the table are computed by summing the respective row or column entries. For instance, in the first column, $n(E) = n(E \cap F) + n(E \cap F^{C})$, since elements in E must belong to either F or F^{C}.

This table can also be constructed using probabilities:

	E	E^{C}	
F	$P(E \cap F)$	$P(E^{C} \cap F)$	$P(F)$
F^{C}	$P(E \cap F^{C})$	$P(E^{C} \cap F^{C})$	$P(F^{C})$
	$P(E)$	$P(E^{C})$	$P(S) = 1$

Note that in this table, the marginal probabilities are again the sum of their respective row or column entries, which can be verified using Exercise 1.2.

Example 1.8.1 *A study was conducted to find the relationship between dog ownership (D) and cat ownership (C) among pet owners in the U.S. In a sample of 300 American pet owners, the researchers found that 219 owned dogs, 12 had neither cats nor dogs, and 72 owned both cats and dogs.*[6]

1. *Construct a contingency table for cat and dog owners.*
2. *Find* $P(C \cap D)$.
3. *Find* $P(C^{\mathsf{c}})$.
4. *Find* $P(D|C^{\mathsf{c}})$.

Solution

1. From the information given in the problem statement:

	D	D^{c}	
C	72		
C^{c}		12	
	219		300

Completing the remaining entries in the table yields:

	D	D^{c}	
C	72	69	141
C^{c}	147	12	159
	219	81	300

2. Using the frequentist interpretation of probability and the contingency table results, we derive:

$$P(C \cap D) = \frac{n(C \cap D)}{n(S)} = \frac{72}{300} = 0.24$$

3.

$$P(C^{\mathsf{c}}) = \frac{n(C^{\mathsf{c}})}{n(S)} = \frac{159}{300} = 0.53$$

4.

$$P(D \cap C^{\mathsf{c}}) = \frac{n(D \cap C^{\mathsf{c}})}{n(S)} = \frac{147}{300} = 0.49$$

$$P(D|C^{\mathsf{c}}) = \frac{P(D \cap C^{\mathsf{c}})}{P(C^{\mathsf{c}})} = \frac{0.49}{0.53} \approx 0.92$$

---

[6] based on data from `https://pewrsr.ch/3JPtwoR`

Hence, it is highly probable that pet owners without cats prefer dogs over other pets.

Another way to calculate $P(D|C^{\mathsf{C}})$ is:

$$P(D|C^{\mathsf{C}}) = \frac{P(D \cap C^{\mathsf{C}})}{P(C^{\mathsf{C}})} = \frac{\frac{n(D\cap C^{\mathsf{C}})}{n(S)}}{\frac{n(C^{\mathsf{C}})}{n(S)}} = \frac{n(D \cap C^{\mathsf{C}})}{n(C^{\mathsf{C}})} = \frac{147}{159} \approx 0.92$$

■

Exercise 1.4 *In the previous exercise, find* $P(C|D^{\mathsf{C}})$*, compare it with* $P(D|C^{\mathsf{C}})$ *and interpret the results.*

1.9

Independent Events

Suppose E and F are two events with positive probabilities. It is possible that prior information that F has occurred does not affect the occurrence of E, or in other words:

$$P(E|F) = P(E)$$

If this is the case, we say E is independent of F. Next, we show that if E is independent of F, F is also independent of E.

By definition, $P(E|F) = \frac{P(E\cap F)}{P(F)}$ which implies $P(E \cap F) = P(E|F)P(F)$, also known as **multiplication rule** (not to be confused with the multiplication principle!). Since E is independent of F, we have $P(E|F) = P(E)$, and consequently:

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

An additional observation from $P(E|F) = P(E)$ is:

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

We therefore define E and F as **independent events** if any of the following equivalent conditions hold:

- $P(E|F) = P(E)$
- $P(F|E) = P(F)$

- $P(E \cap F) = P(E)P(F)$

Note that two disjoint events are not necessarily independent. For E and F to be both independent and disjoint, these conditions should hold:

$$P(E \cap F) = P(E)P(F), E \cap F = \emptyset$$

In other words, at least one of these two events must have a probability of zero.

1.10

Bayes' Theorem

In this section we see an important formula which utilizes conditional probability. Before that, we discuss some important prerequisites.

1.10.1 Partitioning the sample space into mutually exclusive events

Suppose events $E_1, E_2, \ldots, E_k$ with positive probabilities are mutually exclusive and their union is the sample space, S. In other words,

$$E_i \cap E_j = \emptyset$$

for $i \neq j$ where $i, j = 1, 2, \ldots k$, and

$$\bigcup_{i=1}^{k} E_i = S$$

In this case, we say S is partitioned into these events.

Example 1.10.1 *We can partition the sample space in Example 1.6.2 to events "zero heads", "one head", "two heads" and "three heads" as follows:*

$$\{TTT\}, \{HTT, THT, TTH\}, \{HHT, HTH, THH\}, \{HHH\}$$

1.10.2 Partitioning an event into mutually exclusive events

Suppose S is partitioned into events $E_1, E_2, \ldots, E_k$. For any other event F with positive probability we have:

$$F = S \cap F = (\bigcup_{i=1}^{k} E_i) \cap F = \bigcup_{i=1}^{k} (E_i \cap F)$$

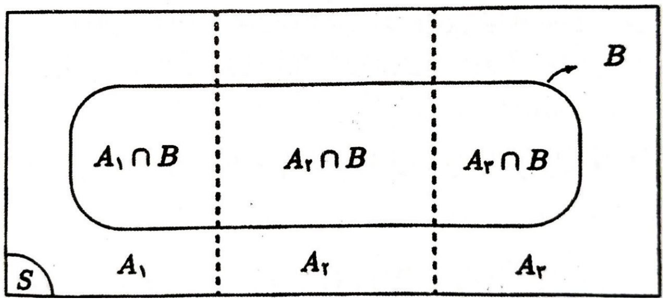


Figure 1.2: Partitioning event B into three events

In this case, we say F is partitioned into events $F \cap E_1, F \cap E_2, \ldots, F \cap E_k$. Note that since $E_1, E_2, \ldots, E_k$ are mutually exclusive, $F \cap E_1, F \cap E_2, \ldots, F \cap E_k$ are also disjoint events. This is also illustrated in Figure 1.2. In this figure, B is partitioned into three events $A_1 \cap B, A_2 \cap B,$ and $A_3 \cap B$.

Example 1.10.2 *Consider an experiment where a card is drawn at random from a deck numbered from 1 to 20. Suppose the sample space S is partitioned into two events:*

1. *E: drawing an even-numbered card*
2. *O: drawing an odd-numbered card*

Then, the event A, "drawing a multiple of 5", can be partitioned into two events:

1. *$A \cap O = \{5, 15\}$ (odd multiples of 5)*
2. *$A \cap E = \{10, 20\}$ (even multiples of 5)*

1.10.3 Law of Total Probability

If F is partitioned into events $E_1, E_2, \ldots, E_n$, then Theorem 1.5.2 yields:

$$P(F) = P\left(\bigcup_{i=1}^{k}(E_i \cap F)\right) = \sum_{i=1}^{k} P(E_i \cap F)$$

Since $P(F|E_i) = \frac{P(F \cap E_i)}{P(E_i)}$, we can alternatively write the above formula as:

$$P(F) = P\left(\bigcup_{i=1}^{k}(F \cap E_i)\right) = \sum_{i=1}^{k} P(F|E_i)P(E_i)$$

which is called **the law of total probability**.

1.10.4 Bayes' Rule

From the definition of conditional probability we know:

$$P(E_j|F) = \frac{P(E_j \cap F)}{P(F)}$$

where $j = 1, 2, \ldots, k$.

Utilizing the law of total probability in the ablove formula yields:

$$P(E_j|F) = \frac{P(F|E_j)P(E_j)}{\sum_{i=1}^{k} P(F|E_i)P(E_i)}$$

This result is known as **Bayes' rule**, which provides a method for computing conditional probabilities when the sample space is partitioned into mutually exclusive events.

Example 1.10.3 *We have three boxes. The first box contains two black marbles and one red marble, the second contains two red marbles and two black marbles, and the third contains three black marbles and two red marbles. Assume the boxes and marbles differ only in their colors and identities. We randomly select one box and then draw one marble from it.*

1. *What is the probability that the drawn marble is black?*
2. *Given that the drawn marble is black, what is the probability that it came from third box?*

Solution A box is chosen and then a marble is drawn from that box. The sample space is $S = \{B_1, B_2, R_1, R_2, R_3, B_3, B_4, B_5, B_6, B_7, R_4, R_5\}$ since no two marbles are identical. S is partitioned into three events depending on the three boxes containing the marbles: A_1, A_2, and A_3, where $A_1 = \{B_1, B_2, R_1\}, A_2 = \{R_2, R_3, B_3, B_4\}$, and $A_3 = \{B_5, B_6, B_7, R_4, R_5\}$. In other words, A_i represents the event that the chosen box is the ith box and so the marble drawn would be an element of A_i. There are no restrictions on choosing any box, and so the best model is a uniform probability model. Thus:

$$P(A_1) = P(A_2) = P(A_3) = \frac{1}{3}$$

Suppose B is the event "the marble drawn from the chosen box is black". Depending on the chosen box, there are three cases to consider:

$$P(B|A_1) = \frac{2}{3}, P(B|A_2) = \frac{2}{4} = \frac{1}{2}, P(B|A_3) = \frac{3}{5}$$

Let us analyze this more thoroughly: for example, suppose the chosen box is the second one. This means that the sample space is reduced to $A_2 = \{R_2, R_3, B_3, B_4\}$, and hence the event "choosing a black marble" is $\{B_3, B_4\}$.

1. By the law of total probability:

$$\begin{aligned} P(B) &= P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + P(B|A_3)P(A_3) \\ &= (\frac{2}{3})(\frac{1}{3}) + (\frac{1}{2})(\frac{1}{3}) + (\frac{3}{5})(\frac{1}{3}) = \frac{53}{90} \end{aligned}$$

2. By the Bayes' rule:

$$P(A_3|B) = \frac{P(B|A_3)P(A_3)}{P(B)} = \frac{(\frac{3}{5})(\frac{1}{3})}{\frac{53}{90}} = \frac{18}{53}$$

■

1.11

Sensitivity, Specificity, Prevalence, and Relative Risk

The importance of Bayes' Theorem is probably best understood through biomedical scenarios. Consider a disease and a test developed for it which determines whether a patient has that disease or not. Suppose a patient does the test, and the test is either positive or negative. But there is uncertainty involved here: the test does not always accurately identify the presence or absence of the disease.

Define T as the event that "the test is positive". So T^C is the event that "the test is negative". We also define D as the event that "the patient has the disease". Now in case $P(T|D)$ and $P(T^C|D^C)$ are given, the probabilty that the patient has the disease if their test is positive can be calculated using the Bayes' rule:

$$P(D|T) = \frac{P(T|D)P(D)}{P(T|D)P(D) + P(T|D^C)P(D^C)}$$

where $P(T|D^C) = 1 - P(T^C|D^C)$ (why?).

$P(T|D)$ is called **sensitivity** of the test, and $P(T^C|D^C)$ the **specificity** of it. In other words, sensitivity is the probability that the test correctly identifies patients who have the disease, and specificity the probability that the test correctly identifies those who do not have the disease.

These terms can be used in a more general binary classification framework and are not restricted to biomedical applications. Consider an algorithm developed to categorize emails as "spam" or "not spam". Within this framework, we must clearly define what constitutes a positive and negative result. For instance, one researcher might classify "spam" as positive and "not spam" as negative, while another could adopt the opposite convention. The emails that are correctly identified as positive are called **true positives**, and the ones that are correctly identified as negative are called **true negatives**. On the other hand, those falsely identified as positives are called **false positives**, and those falsely identified as negatives, **false negatives**. The probability that this algorithm identifies true positives correctly is sensitivity or equivalently true positive rate, and the probability that it identifies true negatives is specificity or equivalently true negative rate.

Example 1.11.1 *Suppose we define "spam" as positive and "non-spam" as negative. An algorithm with 90% sensitivity correctly identifies 90 out of 100 actual spams as spam (true positives), and misclassifies 10 as non-spam (false negatives).*

Example 1.11.2 *Consider an HIV test that classifies HIV-positive cases as positive and HIV-negative cases as negative. With 97% specificity, the test correctly identifies 97 out of 100 truly HIV-negative individuals as negative (true negatives), while incorrectly classifying 3 as positive (false positives).*

Example 1.11.3 *Developers at a food company have created a sentiment analysis algorithm that classifies customer feedback as either "satisfied" (positive) or "dissatisfied" (negative). From a test dataset of 500 comments, 200 were known to be from satisfied customers and 300 from dissatisfied customers. The algorithm classified 150 of the true satisfied comments correctly as positive, but incorrectly labeled 20 of the dissatisfied comments as positive. Using these results, calculate the algorithm's sensitivity and specificity.*

Solution We first create a contingency table based on the data:

	H	H^{c}	
P	150	20	170
P^{c}	50	280	330
	200	300	500

where H is the event "customer is satisfied" and P the event "comment is labeled as positive". Sensitivity is calculated as:

$$P(P|H) = \frac{P(P \cap H)}{P(H)} = \frac{\frac{150}{500}}{\frac{200}{500}} = 0.75$$

and specificity as:

$$P(P^{\mathsf{c}}|H^{\mathsf{c}}) = \frac{P(P^{\mathsf{c}} \cap H^{\mathsf{c}})}{P(H^{\mathsf{c}})} = \frac{\frac{280}{500}}{\frac{300}{500}} \approx 0.93$$

■

Sometimes we are interested in the prportion of a population that has a characteristic. **Prevalence** is defined as:

$$\frac{\#\text{ of people with characteristic}}{\text{total }\#\text{ of people}}$$

Example 1.11.4 *A hospital evaluates a new rapid diagnostic test for influenza. During a flu season trial, 500 patients with flu-like symptoms were tested using both the rapid test and a gold-standard PCR test. The results showed that 120 patients tested positive on both tests, while 30 patients tested positive on the rapid test but negative on PCR. Among those who tested negative on the rapid test, 50 were actually PCR-positive.*

1. *Construct the contingency table summarizing these results.*
2. *Calculate the prevalence of influenza in this patient population based on the gold-standard PCR test.*

Solution

- We define the event R as "the patient tests positive on the rapid test", and G as "the patient tests positive on the PCR test".

	R	R^C	
G	120	50	170
G^C	30	300	330
	150	350	500

- From the contingency table above we infer that among these 500 tested patients, 170 tested positive based on the gold standard PCR test. So the prevalence measure is $\frac{170}{500} = 0.34$.

■

Another important concept in biomedical applications is **relative risk (RR)**, which is the probability of an event occurring in one group relative to the probability of the event occurring in another group:

$$RR = \frac{\hat{p}_1}{\hat{p}_2}$$

where $\hat{p}_1$ is the probability of the event occurring in one group and $\hat{p}_2$ probability of it happening in another.

Three cases may happen:

- $RR = 1$: the event occurs with equal probability in each group
- $RR > 1$: the event is more likely to happen in the first group
- $RR < 1$: the event is more likely to happen in the second group

Example 1.11.5 *A public health study investigated the association between dietary salt intake and hypertension in a population of 1,000 adults. Researchers classified participants into two groups based on their daily salt consumption: 500 people reported high salt intake (> 5 grams per day) and 500 reported moderate salt intake (≤ 5 grams per day). After follow-up, 150 cases of hypertension were identified in the high-salt group, while only 50 cases occurred in the moderate-salt group.*

1. *Organize these data into a contingency table comparing salt intake (high vs moderate) with hypertension status (present vs absent).*
2. *Calculate the prevalence of hypertension in this sample.*
3. *Compute relative risk of hypertension between high salt intake and moderate intake samples.*

Solution

1. We define the event H as "the participant has hypertension", and I as "the participant has high salt intake".

	H	H^c	
I	150	350	500
I^c	50	450	500
	200	800	1000

2. $\frac{200}{1000} = 0.2$
3. Suppose the first group is the high salt intake cohort and the second one comprises those with moderate intake. Thus:

$$\hat{p}_1 = \frac{150}{500} = 0.3, \hat{p}_2 = \frac{50}{500} = 0.1$$
$$RR = \frac{\hat{p}_1}{\hat{p}_2} = \frac{0.3}{0.1} = 3$$

   Since $RR > 1$, the first group, which is the high salt intake cohort is more likey to have hypertension. Note that we are not inferring high salt intake causes hypertension! This is the topic of a future chapter.

■

Glossary

2. Probability and Accuracy

2.1
Probability

2.2
Conditional Probability & Bayes' Theorem

2.3
Prevalence, Relative Risk, Sensitivity, and Specificity

Glossary

3. Descriptive Statistics

3.1
Introduction to Data

3.2
Exploratory Data Analysis

Glossary

4. Probability Distributions

4.1
Random Variables

4.2
Binomial Distribution

4.3
Normal Distribution

4.4
Sampling Distribution and Central Limit Theorem

Glossary

5. Estimation

5.1

Point Estimation

5.2

Confidence Interval

Glossary

6. Estimation

6.1

Hypothesis Testing for One Sample

6.2

Hypothesis Testing for Two Variances

6.3

Hypothesis for Two Means

6.4

Hypothesis for Two Proportions

Glossary

7. Hypothesis Testing

Glossary

- **anecdotal evidence**: Evidence, often personal, that is collected casually rather than by a well-designed study.
- **population**: A group we are interested in studying. "Population" often refers to a group of people, but the term is used for other subjects, too.
- **cross-sectional study**: A study that collects data about a population at a particular point in time.
- **cycle**: In a repeated cross-sectional study, each repetition of the study is called a cycle.
- **longitudinal study**: A study that follows a population over time, collecting data from the same group repeatedly.
- **record**: In a dataset, a collection of information about a single person or other subject.
- **respondent**: A person who responds to a survey.
- **sample**: The subset of a population used to collect data.
- **representative**: A sample is representative if every member of the population has the same chance of being in the sample.
- **oversampling**: The technique of increasing the representation of a sub-population in order to avoid errors due to small sample sizes.
- **raw data**: Values collected and recorded with little or no checking, calculation or interpretation.
- **recode**: A value that is generated by calculation and other logic applied to raw data.
- **data cleaning**: Processes that include validating data, identifying errors, translating between data types and representations, etc.

8.

Glossary

- **linear fit**: a line intended to model the relationship between variables.
- **least squares fit**: A model of a dataset that minimizes the sum of squares of the residuals.
- **residual**: The deviation of an actual value from a model.
- **goodness of fit**: A measure of how well a model fits data.
- **coefficient of determination**: A statistic intended to quantify goodness of fit.
- **sampling weight**: A value associated with an observation in a sample that indicates what part of the population it represents.

9. Probability and Accuracy

Glossary

- **regression**: One of several related processes for estimating parameters that fit a model to data.
- **dependent variables**: The variables in a regression model we would like to predict. Also known as endogenous variables.
- **explanatory variables**: The variables used to predict or explain the dependent variables. Also known as independent, or exogenous, variables.
- **simple regression**: A regression with only one dependent and one explanatory variable.
- **multiple regression**: A regression with multiple explanatory variables, but only one dependent variable.
- **linear regression**: A regression based on a linear model.
- **ordinary least squares**: A linear regression that estimates parameters by minimizing the squared error of the residuals.
- **spurious relationship**: A relationship between two variables that is caused by a statistical artifact or a factor, not included in the model, that is related to both variables.
- **control variable**: A variable included in a regression to eliminate or "control for" a spurious relationship.
- **proxy variable**: A variable that contributes information to a regression model indirectly because of a relationship with another factor, so it acts as a proxy for that factor.
- **categorical variable**: A variable that can have one of a discrete set of unordered values.
- **join**: An operation that combines data from two DataFrames using a key to match up rows in the two frames.

- **data mining**: An approach to finding relationships between variables by testing a large number of models.
- **logistic regression**: A form of regression used when the dependent variable is boolean.
- **Poisson regression**: A form of regression used when the dependent variable is a non-negative integer, usually a count.
- **odds**: An alternative way of representing a probability, p, as the ratio of the probability and its complement, $p/(1-p)$.

10.

Glossary

- **time series**: A dataset where each value is associated with a timestamp, often a series of measurements and the times they were collected.
- **window**: A sequence of consecutive values in a time series, often used to compute a moving average.
- **moving average**: One of several statistics intended to estimate the underlying trend in a time series by computing averages (of some kind) for a series of overlapping windows.
- **rolling mean**: A moving average based on the mean value in each window.
- **exponentially-weighted moving average (EWMA)**: A moving average based on a weighted mean that gives the highest weight to the most recent values, and exponentially decreasing weights to earlier values.
- **span**: A parameter of EWMA that determines how quickly the weights decrease.
- **serial correlation**: Correlation between a time series and a shifted or lagged version of itself.
- **lag**: The size of the shift in a serial correlation or autocorrelation.
- **autocorrelation**: A more general term for a serial correlation with any amount of lag.
- **autocorrelation function**: A function that maps from lag to serial correlation.
- **stationary**: A model is stationary if the parameters and the distribution of residuals does not change over time.

11.

Glossary

- **survival analysis**: A set of methods for describing and predicting lifetimes, or more generally time until an event occurs.
- **survival curve**: A function that maps from a time, t, to the probability of surviving past t.
- **hazard function**: A function that maps from t to the fraction of people alive until t who die at t.
- **Kaplan-Meier estimation**: An algorithm for estimating hazard and survival functions.
- **cohort**: a group of subjects defined by an event, like date of birth, in a particular interval of time.
- **cohort effect**: a difference between cohorts.
- **NBUE**: A property of expected remaining lifetime, “New better than used in expectation.”
- **UBNE**: A property of expected remaining lifetime, “Used better than new in expectation.”

12.

Glossary

- **distribution**: The values that appear in a sample and the frequency of each.
- **histogram**: A mapping from values to frequencies, or a graph that shows this mapping.
- **frequency**: The number of times a value appears in a sample.
- **mode**: The most frequent value in a sample, or one of the most frequent values.
- **normal distribution**: An idealization of a bell-shaped distribution; also known as a Gaussian distribution.
- **uniform distribution**: A distribution in which all values have the same frequency.
- **tail**: The part of a distribution at the high and low extremes.
- **central tendency**: A characteristic of a sample or population; intuitively, it is an average or typical value.
- **outlier**: A value far from the central tendency.
- **spread**: A measure of how spread out the values in a distribution are.
- **summary statistic**: A statistic that quantifies some aspect of a distribution, like central tendency or spread.
- **variance**: A summary statistic often used to quantify spread.
- **standard deviation**: The square root of variance, also used as a measure of spread.
- **effect size**: A summary statistic intended to quantify the size of an effect like a difference between groups.
- **clinically significant**: A result, like a difference between groups, that is relevant in practice.

13.

Glossary

- **Probability mass function (PMF)**: a representation of a distribution as a function that maps from values to probabilities.
- **probability**: A frequency expressed as a fraction of the sample size.
- **normalization**: The process of dividing a frequency by a sample size to get a probability.
- **index**: In a pandas DataFrame, the index is a special column that contains the row labels.

14.

Glossary

- **percentile rank**: The percentage of values in a distribution that are less than or equal to a given value.
- **percentile**: The value associated with a given percentile rank.
- **cumulative distribution function (CDF)**: A function that maps from values to their cumulative probabilities. $CDF(x)$ is the fraction of the sample less than or equal to x.
- **inverse CDF**: A function that maps from a cumulative probability, p, to the corresponding value.
- **median**: The 50th percentile, often used as a measure of central tendency.
- **interquartile range**: The difference between the 75th and 25th percentiles, used as a measure of spread.
- **quantile**: A sequence of values that correspond to equally spaced percentile ranks; for example, the quartiles of a distribution are the 25th, 50th and 75th percentiles.
- **replacement**: A property of a sampling process. “With replacement” means that the same value can be chosen more than once; “without replacement” means that once a value is chosen, it is removed from the population.

15.

Glossary

- **empirical distribution**: The distribution of values in a sample.
- **analytic distribution**: A distribution whose CDF is an analytic function.
- **model**: A useful simplification. Analytic distributions are often good models of more complex empirical distributions.
- **interarrival time**: The elapsed time between two events.
- **complementary CDF**: A function that maps from a value, x, to the fraction of values that exceed x, which is $1 - CDF(x)$.
- **standard normal distribution**: The normal distribution with mean 0 and standard deviation 1.
- **normal probability plot**: A plot of the values in a sample versus random values from a standard normal distribution.

16.

Glossary

- **Probability density function (PDF)**: The derivative of a continuous CDF, a function that maps a value to its probability density.
- **Probability density**: A quantity that can be integrated over a range of values to yield a probability. If the values are in units of cm, for example, probability density is in units of probability per cm.
- **Kernel density estimation (KDE)**: An algorithm that estimates a PDF based on a sample.
- **discretize**: To approximate a continuous function or distribution with a discrete function. The opposite of smoothing.
- **raw moment**: A statistic based on the sum of data raised to a power.
- **central moment**: A statistic based on deviation from the mean, raised to a power.
- **standardized moment**: A ratio of moments that has no units.
- **skewness**: A measure of how asymmetric a distribution is.
- **sample skewness**: A moment-based statistic intended to quantify the skewness of a distribution.
- **Pearson's median skewness coefficient**: A statistic intended to quantify the skewness of a distribution based on the median, mean, and standard deviation.
- **robust**: A statistic is robust if it is relatively immune to the effect of outliers.

17. Probability and Accuracy

17.1

Probability

17.2

Conditional Probability & Bayes’ Theorem

17.3

Prevalence, Relative Risk, Sensitivity, and Specificity

Glossary

- **scatter plot**: A visualization of the relationship between two variables, showing one point for each row of data.
- **jitter**: Random noise added to data for purposes of visualization.
- **saturation**: Loss of information when multiple points are plotted on top of each other.
- **correlation**: A statistic that measures the strength of the relationship between two variables.
- **standardize**: To transform a set of values so that their mean is 0 and their variance is 1.
- **standard score**: A value that has been standardized so that it is expressed in standard deviations from the mean.
- **covariance**: A measure of the tendency of two variables to vary together.

- **rank**: The index where an element appears in a sorted list.
- **randomized controlled trial**: An experimental design in which subjects are divided into groups at random, and different groups are given different treatments.
- **treatment group**: A group in a controlled trial that receives some kind of intervention.
- **control group**: A group in a controlled trial that receives no treatment, or a treatment whose effect is known.
- **natural experiment**: An experimental design that takes advantage of a natural division of subjects into groups in ways that are at least approximately random.

18. Descriptive Statistics

18.1

Introduction to Data

18.2

Exploratory Data Analysis

Glossary

- **estimation**: The process of inferring the parameters of a distribution from a sample.
- **estimator**: A statistic used to estimate a parameter.
- **mean squared error (MSE)**: A measure of estimation error.
- **root mean squared error (RMSE)**: The square root of MSE, a more meaningful representation of typical error magnitude.
- **maximum likelihood estimator (MLE)**: An estimator that computes the point estimate most likely to be correct.
- **bias (of an estimator)**: The tendency of an estimator to be above or below the actual value of the parameter, when averaged over repeated experiments.
- **sampling error**: Error in an estimate due to the limited size of the sample and variation due to chance.
- **sampling bias**: Error in an estimate due to a sampling process that is not representative of the population.

- **measurement error**: Error in an estimate due to inaccuracy collecting or recording data.
- **sampling distribution**: The distribution of a statistic if an experiment is repeated many times.
- **standard error**: The RMSE of an estimate, which quantifies variability due to sampling error (but not other sources of error).
- **confidence interval**: An interval that represents the expected range of an estimator if an experiment is repeated many times.

19. Probability Distributions

19.1
Random Variables

19.2
Binomial Distribution

19.3
Normal Distribution

19.4
Sampling Distribution and Central Limit Theorem

Glossary

- **hypothesis testing**: The process of determining whether an apparent effect is statistically significant.
- **test statistic**: A statistic used to quantify an effect size.
- **null hypothesis**: A model of a system based on the assumption that an apparent effect is due to chance.
- **p-value**: The probability that an effect could occur by chance.

- **statistically significant**: An effect is statistically significant if it is unlikely to occur by chance.
- **permutation test**: A way to compute p-values by generating permutations of an observed dataset.
- **resampling test**: A way to compute p-values by generating samples, with replacement, from an observed dataset.
- **two-sided test**: A test that asks, “What is the chance of an effect as big as the observed effect, positive or negative?”
- **one-sided test**: A test that asks, “What is the chance of an effect as big as the observed effect, and with the same sign?”
- **chi-squared test**: A test that uses the chi-squared statistic as the test statistic.
- **false positive**: The conclusion that an effect is real when it is not.
- **false negative**: The conclusion that an effect is due to chance when it is not.
- **power**: The probability of a positive test if the null hypothesis is false.