2 Exploring probability and counting

This chapter covers

- Basic probabilities
- Counting rules
- Combinations versus permutations
- Continuous random variables
- Discrete random variables

Our combined purpose in this chapter and chapter 3 is to explore the fascinating world of probabilities from several different angles—from the very basics to more advanced topics like conditional probabilities. Probabilities are the cornerstone of statistical analysis and many other quantitative techniques, so it makes perfect sense to start here and paint the subject of probabilities with a broad brush before tackling more specific topics in subsequent chapters.

Probability theory is not just some mathematical abstraction; on the contrary, it is a powerful tool for making informed decisions across almost any professional field—and in our personal lives. By understanding probabilities, we can quantify uncertainty, assess risks, and make informed predictions from data. Mastering probability concepts not only establishes a solid foundation from which to approach the techniques covered in the rest of this book but also opens up a world of possibilities around analyzing data, making predictions, and drawing meaningful and actionable insights.

Our journey in chapter 2 starts small with the basics of probabilities and odds and finishes big with conditional probability calculations. In between, we'll examine fundamental counting rules, evaluate the distinctions between combinations and permutations, and explore the properties of continuous and discrete random variables. Then, in chapter 3, we'll inspect four probability distributions—normal, binomial, uniform, and Poisson—that are foundational in statistical and quantitative analysis, and review the fundamental principles underlying probability computations. Let's begin by getting grounded in the basics.

2.1 Basic probabilities

In everyday life, numerous events unfold as a result of randomness or chance, ranging from the mundane to the consequential. Consider the simple act of flipping a coin, where the outcome—heads or tails—is beyond our influence. Similarly, when rolling a pair of dice, the probability of rolling a 7—an outcome more likely than any other possible result—is still subject to chance. Or think about selecting a single card from a standard deck of 52 cards, which are equally divided into four suits and further split into three types of face cards and 10 numbered cards; the probabilities of selecting *any* face card or *any* numbered card still exist in a state of unpredictability.

However, when we know the total number of possible outcomes and can identify how many of those lead to a successful event, we can calculate the probability by comparing the number of successful outcomes to the total possible outcomes. This enables us to make probability-based decisions grounded in the ratio of favorable outcomes to all potential outcomes. In fact, when we think about probabilities, we are really focusing on the likelihood of a specific event occurring, such as getting heads on a coin flip after calling heads. Therefore, the probability of a successful event can be expressed as follows:

$$\text{Probability(Event)} = \frac{\text{total number of potential outcomes that are successful}}{\text{total number of possible outcomes}}$$

or more simply

$$\text{Probability(Event)} = \frac{\text{number of successes}}{\text{number of outcomes}}$$

Two prerequisites are absolutely critical: we need to get the denominator right, and there must be agreement about what qualifies as a success. Getting the denominator right is easy when flipping a two-sided coin. But it's not as easy when rolling two six-sided dice or selecting cards from a deck without replacement. Likewise, agreement about what constitutes a success is easy when we call heads on a coin flip and get heads, but we also might want to qualify a product malfunction as a “success” if such malfunctions are rare and therefore easy to track.

Let’s explore some simple and straightforward probabilities together. When we flip a fair coin with two sides, there are two potential outcomes, heads or tails, that are equally likely. If we call heads, and therefore heads is the successful outcome, the probability of success is

$$\frac{1}{2}$$

When we roll two six-sided dice, there are $6 \times 6 = 36$ possible outcomes. There’s only one way of getting double 6s, if that’s what we’re hoping for; so, the probability of success is

$$\frac{1}{36}$$

And when selecting a card from a standard deck of 52 cards, the probability of selecting a face card—that is, a King, Queen, or Jack—regardless of suit is

$$\frac{12}{52} = \frac{6}{26} = \frac{3}{13}$$

Let’s now assume we’ve flipped a coin and it came up heads. The probability that the same coin will come up heads on the next flip is

$$\frac{1}{2}$$

The probability is the same because these are independent events. Just because the first flip came up heads doesn’t increase the likelihood that the next flip will result in the other potential outcome. Or consider 35 rolls of two six-sided dice, none of which resulted in double 6s. The probability that the next roll will result in a pair of 6s is still

$$\frac{1}{36}$$

In fact, it will always be the same probability—the number of potential outcomes doesn’t change, nor does the number of ways to get a successful outcome change.

Or consider selecting a face card. As long as each selected card is then returned to the deck, the probability of getting a face card will always be equal to

$$\frac{3}{13}$$

The probability of a successful event is independent of the outcomes of previous like events. Although we should expect, for instance, the counts of heads and tails to converge after hundreds, or thousands, of coin flips—this is the essence of the law of large numbers, by the way—this phenomenon is completely separate and unrelated to the next flip of the coin.

So far, we've been discussing *theoretical probabilities*. That's because flipping coins, rolling dice, and selecting cards are based on a theoretical or assumed understanding of equally likely outcomes in a fixed and controlled environment. But there are other types of probabilities, which we'll discuss next.

2.1.1 Probability types

We have covered one type of probability already: what we call *theoretical probabilities*. Let's consider two other types of probability: empirical and subjective.

Empirical probabilities are derived from trials or real-world observations. The probability of success is therefore expressed this way:

$$\text{Probability(Event)} = \frac{\text{number of successes observed}}{\text{number of observations made}}$$

For instance, consider the probability of rainfall in Cincinnati, Ohio. Suppose meteorologists have been collecting data on rainfall for several years. They've recorded whether it rains or not each day and have otherwise kept a running tally of rainy days versus the number of days overall. Empirical probability in this context would involve calculating the probability of rainfall based on the observed data. So, if the denominator equals 365 days and the numerator equals 75 (because it's previously rained 75 times over that span), then the empirical probability of rainfall on any given day would be calculated as

$$\text{Probability(rainfall)} = \frac{\text{number of rainy days}}{\text{total number of days}}$$

or

$$\text{Probability(rainfall)} = \frac{75}{365}$$

And then there's subjective probability. *Subjective probability* is based on personal judgment, opinions, or beliefs rather than on observed data or mathematical calculations; it can therefore vary from person to person and may be influenced by individual experiences, biases, or perceptions. For example, a person's subjective probability of winning a chess match may be higher if they feel confident in their skill and experience, even if the theoretical probability of winning is lower due to objective factors.

2.1.2 Converting and measuring probabilities

Probabilities are frequently converted to percentages and odds and then presented in these terms. Let's first demonstrate how probabilities are converted to percentages by using Python as a calculator.

The following snippet of code calculates the probability of getting heads in a single coin flip, converts the probability to a percentage by multiplying the result by `100`, and prints the result by combining a character string with the percentage probability:

```
>>> probability_heads = 1 / 2
>>> probability_heads_percent = probability_heads * 100
>>> print(f'The probability of success equals: '
      f'{probability_heads_percent}%')
The probability of success equals: 50.0%
```

NOTE A few notes about this snippet of code and many others to follow: first, lines of Python code are always preceded by `>>>`, whereas the results copied and pasted from the Python Console are not. Second, the backslash character (\) is frequently used to cleanly split long lines of Python code into multiple lines. And third, single or double quotation marks are equally acceptable in Python, but consistency is required for each character string, or Python will throw an error.

The following snippets of code perform similar operations with respect to getting double 6s or selecting a face card:

```
>>> probability_double_sixes = 1 / 36
>>> probability_double_sixes_percent = probability_double_sixes * 100
>>> print(f'The probability of success equals: '
>>>       f'{probability_double_sixes_percent}%')
The probability of success equals: 2.7777777777777777%
```

```
>>> probability_face_card = 12 / 52
>>> probability_face_card_percent = probability_face_card * 100
>>> print(f'The probability of success equals: '
>>>       f'{probability_face_card_percent}%')
The probability of success equals: 23.076923076923077%
```

There is thus a 50% chance of getting heads on a coin flip, an almost 3% chance of getting double 6s from a pair of dice, and a 23% chance of selecting a face card from a standard deck of 52 cards.

Odds, meanwhile, are expressed as the ratio between successes and failures, whereas probability represents the ratio of successes to total possible outcomes:

$$\text{Odds} = \frac{\text{number of potential successes}}{\text{number of potential failures}}$$

Although related, odds and probabilities are clearly not the same and should not be used interchangeably.

The odds of selecting a face card are

$$\text{Odds(face card)} = \frac{\text{number of potential successes}}{\text{number of potential failures}}$$

or

$$\text{Odds(face card)} = \frac{12}{40}$$

The numerator and denominator, when added together, equal the total number of possible outcomes. Simplifying the previous fraction, we get

$$\text{Odds(face card)} = \frac{3}{10}$$

or

$$\text{Odds(face card)} = 0.3$$

Alternatively, rather than assigning the number of successes to the numerator and the number of failures to the denominator, we can instead assign the probabilities of success and failure, which must sum to 1, to get the same result:

$$\text{Odds(face card)} = \frac{0.23}{0.77}$$

or

$$\text{Odds(face card)} = 0.3$$

And incidentally, we can go in reverse; that is, we can derive the probability of selecting a face card from the odds of doing the same. We simply insert the odds into the following formula:

$$\text{Probability(Event)} = \frac{\text{Odds}}{\text{Odds} + 1}$$

So that

$$\text{Probability(face card)} = \frac{0.3}{0.3 + 1}$$

or

$$\text{Probability(face card)} = \frac{0.3}{1.3}$$

All this returns us to a 23% probability of selecting a face card from a standard deck of 52 cards when the fractional result is converted to a percentage.

We previously mentioned that it can sometimes be challenging to compute the number of potential outcomes. Such challenges can be resolved by understanding fundamental counting rules, the differences between combinations and permutations, and the concepts of replacement and without replacement. We'll look at this topic next.

2.2 Counting rules

In probability and combinatorics (the study of counting, arrangement, and combination of objects), understanding how to count the number of possible outcomes is paramount. A pair of fundamental principles are our guides: the multiplication rule and the addition rule. These two rules are the very foundation for calculating the total number of outcomes when dealing with multiple events or scenarios. The multiplication rule is a method for determining the total number of potential outcomes when events are simultaneous or sequential, and the addition rule is a method for determining the same result when events are mutually exclusive. Together, they are powerful tools for analyzing various probability scenarios and uncovering the myriad ways in which events can unfold.

2.2.1 Multiplication rule

The multiplication rule states that if there are i choices to be made, with n_1 possibilities for the first choice, n_2 possibilities for the second choice, and so forth, then the aggregate number of potential outcomes equals the product of the individual choices, denoted as $n_1 \times n_2 \times \ldots n_i$. To illustrate, when rolling a pair of six-sided dice, we used the multiplication rule (rolling two dice consists of two simultaneous events, after all) to compute the number of possible outcomes, arriving at 36 by multiplying 6 by itself. It's easy to envision so many possible outcomes when you realize that getting 3 on the first die and 2 on the second die is not the same outcome as getting 2 on the first die and 3 on the second. Let's explore a pair of other examples where the multiplication rule applies:

- If a license plate contains three letters followed by three numerals (e.g., ABC123), applying the multiplication rule returns the total number of possible outcomes. There are 26 letters in the alphabet and 10 single-digit numerals; therefore, the number of possible outcomes equals $26 \times 26 \times 26 \times 10 \times 10 \times 10$, or 17,576,000 unique license plates.
- Similarly, in the case of a briefcase featuring a three-digit lock, where each digit is any number between 0 and 9, the total number of possible outcomes amounts to $10 \times 10 \times 10$, or 1,000 unique outcomes.

The multiplication rule applies when the outcomes of two or more events are independent of each other, meaning that the occurrence of one event does not influence the occurrence of the other event(s). It allows us to calculate the total number of outcomes for the *combined* events by multiplying the number of outcomes for each individual event.

2.2.2 Addition rule

In situations where events are mutually exclusive, the addition rule is applied to calculate the probability of either event occurring. The addition rule states that the total probability of one or another of the mutually exclusive events occurring is the sum of their individual probabilities. It's important to note that this rule applies specifically to events that cannot happen simultaneously.

For example, consider a scenario where you have the option to either flip a coin or roll a die. The outcomes of these events are mutually exclusive because you can only do one of the actions at a time. If you choose to flip the coin, there are two possible outcomes (heads or tails), and if you choose to roll the die, there are six possible outcomes. According to the addition rule, the total number of potential outcomes is $2 + 6 = 8$, assuming you choose to perform only one of these actions.

However, if the events are independent (i.e., the outcome of one event does not affect the outcome of another), the addition rule does not apply. Instead, the multiplication rule is used to calculate the total number of possible outcomes. For instance, if you were to flip a coin, roll a die, and select a card from a standard deck, the number of possible outcomes would be calculated as $2 \times 6 \times 52 = 624$, reflecting all possible combinations of these independent events.

Here are some additional examples to clarify the application of the addition rule:

- If you choose to read 1 of 12 classic novels or 1 of 14 data science manuals (but not both), the number of possible outcomes is $12 + 14 = 26$.
- If you have three interstate routes and two backroad routes to travel from Cincinnati to Nashville, and you can only choose one of these routes, the total number of possible routes is $3 + 2 = 5$.

It is essential to distinguish between when to apply the addition rule versus the multiplication rule, as each rule applies to different types of events. The addition rule is used when calculating the total number of outcomes for mutually exclusive events, and the multiplication rule is used when dealing with independent events. Understanding the correct application of these rules is crucial for accurately computing probabilities.

2.2.3 Combinations and permutations

We've been chewing on a mix of combinations and permutations. In spite of the fact that *combination* is frequently used as a euphemism for combinations and permutations, the two are actually quite different; and because they are different, the methods by which they are mathematically derived are unalike. And then there's *replacement* and *without replacement*, both of which apply to combinations and permutations, which further complicates matters.

We'll remove the haze, but for now, bear in mind that when the order doesn't matter, it's a combination; but when the order does matter, it's a permutation. When a public address announcer introduces the five starting players before a basketball game, that's a combination, because the sequence of introductions is immaterial; the starting lineup remains unchanged regardless of the order in which the players were introduced. Conversely, consider the three-digit lock on a briefcase, such as 3-7-4. This is a permutation because the specific sequence matters significantly: only 3-7-4 will unlock the briefcase, whereas other arrangements of the same numerals won't work.

PERMUTATIONS WITH REPLACEMENT

Permutations with replacement are relatively simple and straightforward, at least mathematically. A permutation with replacement involves selecting items from a set in a specific order and, after each selection, returning the chosen item to the set before the next selection is made, thereby making the chosen item eligible to be selected again.

For instance, if the three-digit lock on a briefcase can be set as 3-3-4, that's a permutation with replacement because the order matters and reusing digits is allowed. And because the digits are immediately replaced once they're used, the number of choices remains unchanged throughout; it's not decremented by one immediately following each selection. Decrementing by one occurs when replacement is not allowed, meaning each choice reduces the number of available options.

Setting a three-digit lock on a briefcase involves selecting a sequence of single-digit numerals to secure the contents. The process is akin to selecting a password for access, where the order of the numerals matters. When setting the first digit of the lock, we have 10 numerals, 0 through 9, from which to choose. Because the digits can be reused, we then have the same 10 choices when setting the second digit and again when setting the third and final digit.

Thus, the formula to get the number of potential outcomes when the order matters and replacement is allowed is simply

$$n^r$$

where

- n is the number of distinct items (*events* or *choices* could be substituted for *items*).
- r is the number of selections made.
- n^r represents the total number of permutations allowing for replacement after each selection. It is a "take" on the multiplication rule in that both concepts involve the number of outcomes from independent, yet simultaneous or sequential, events.

So, there are 10^3 or $10 \times 10 \times 10 = 1{,}000$ unique outcomes, just as we previously mentioned.

To demonstrate how to calculate the number of permutations with replacement in Python, let's say `n` equals `5` and `r` equals `3`, which is to say we have five choices for each selection and three choices to be made:

```
>>> n = 5
>>> r = 3
```

In Python, one method of raising a number, like `n`, to the power of another, such as `r`, is to call the asterisk (`**`) twice, like so:

```
>>> num_permutations = n ** r
>>> print(f'Number of permutations with replacement: {num_permutations}')
Number of permutations with replacement: 125
```

Alternatively, we can pass `n` and `r` to the `pow()` method to get the same result; `pow()` raises the first argument it takes to the power of the second argument:

```
>>> num_permutations = pow(n, r)
>>> print(f'Number of permutations with replacement: {num_permutations}')
Number of permutations with replacement: 125
```

That was simple and straightforward enough, but permutations are more complicated when replacement is not allowed.

Permutations without replacement

A permutation without replacement means that although the order remains relevant, we have to reduce the number of available choices for each successive selection. Let's say we have five choices for the first selection and a total of three choices to be made. So, we again set `n` to equal `5` and `r` to equal `3`:

```
>>> n = 5
>>> r = 3
```

But this time, we apply the factorial function (!) to `n` and `r` to get the total number of potential outcomes. When we apply the factorial function, we are merely multiplying a series of incrementally descending natural numbers, with the effect of successively decrementing available options. Here are a couple of examples:

- $5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$
- $3! = 3 \times 2 \times 1 = 6$

The formula to calculate the number of potential outcomes when the order matters and replacement is not allowed is given by

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

This formula is derived from the concept that for each position in a sequence, you have one fewer option as each item is used up. Initially, there are n choices for the first position, $n-1$ choices for the second, and so on, until you've made r selections. The factorial in the denominator, $(n-r)!$, accounts for the reduction in available choices as each item is selected, reflecting the fact that replacement is not allowed.

So, when `n` is 5 and `r` is 3, the number of potential outcomes is equal to

$$\frac{5!}{(5-3)!}$$

or

$$\frac{120}{12}$$

which, of course, is equal to 60. So, if any three runners out of five can qualify for a medal, there are 60 first-, second-, and third-place permutations possible.

In Python, we call the `factorial()` method from the `math` library to get the same result:

```
>>> import math
>>> permutations = math.factorial(n) / math.factorial(n - r)
>>> print(f'Number of permutations without replacement: '
>>>       f'{permutations}')
Number of permutations without replacement: 60
```

Or we can instead pass `n` and `r` to the `math.perm()` method, which inserts the assignments for `n` and `r` and then runs the permutations without replacement formula:

```
>>> permutations = math.perm(n, r)
>>> print(f'Number of permutations without replacement: '
>>>       f'{permutations}')
Number of permutations without replacement: 60
```

We get the same result, incidentally, by simply multiplying the three digits, $5 \times 4 \times 3$, that are greater than the denominator. This method offers a quick and intuitive mathematical shortcut to calculate the number of permutations without replacement, bypassing the need to apply the full factorial formula while still arriving at the correct answer. It works because $n! \,/\, (n - r)!$ effectively *removes* the unnecessary tail end of the factorial—that is, everything from $(n - r)!$ downward—leaving just the first r descending terms, which is exactly what you get by multiplying $n \times (n - 1) \times \ldots \times (n - r + 1)$. It's a handy trick that underscores the elegance and efficiency of mathematical thinking.

COMBINATIONS WITHOUT REPLACEMENT

Most lotteries are examples of combinations without replacement—numbers are drawn one at a time (and can't be drawn again), but like the introduction of players at the start of a basketball game, the order in which the numbers are selected has no meaning. Maybe the easiest and most logical way to think about combinations without replacement is to first think about permutations without replacement and then insert an adjustment to eliminate the significance of order.

Let's walk through this together, where n (the number of distinct items) once more equals 5 and r (the number of items to be selected) again equals 3. We already know, from our prior example, that the number of permutations without replacement equals 60. However, when the order no longer matters, the number of distinct outcomes is significantly reduced. For example, the sequences 123, 132, 213, 231, 312, and 321 are all considered different when the order matters; but when the order doesn't matter, these six sequences are treated as identical, leaving only one relevant outcome. To find the number of ways three digits can be sequenced when the order matters, we apply the factorial function, like so:

$$3! = 3 \times 2 \times 1 = 6$$

This means permutations without replacement have six times as many possible outcomes as combinations without replacement when n and r equal 5 and 3, respectively.

Consequently, we adjust the permutations without replacement formula to reduce it by an order of magnitude equal to the number that was just calculated, because we no longer care about the order. Thus, the number of combinations without replacement can be derived by plugging the values for n and r into the following equation:

$$\frac{n!}{(n-r)!} \times \frac{1}{r!} = \frac{n!}{r!\,(n-r)!}$$

or simply

$$\frac{n!}{r!\,(n-r)!}$$

or

$$\frac{5!}{3!\,(5-3)!}$$

or

$$\frac{120}{12}$$

This, of course, equals 10. So when n is 5 and r is 3, the number of permutations without replacement is, indeed, exactly six times more than the number of combinations without replacement. Another, much more interesting, method of getting the same result is to use Pascal's Triangle.

Pascal's triangle

Pascal's triangle is a geometric arrangement of numbers in a triangular shape. The triangle is named after the French mathematician Blaise Pascal, although it was known long before his time. It's commonly used in algebra, probability theory, and combinatorics. Here's how it works.

The triangle is built from the top down: it starts with the single number 1 at the top, and each row below it is constructed by adding the two numbers immediately above it. So, for example, the fourth row in the triangle contains two instances of the number 3—the first instance is obtained by adding the 1 and the 2 immediately above it, and the second instance is obtained by adding the same 2 and a different 1 just above it.

It turns out that each row begins and ends with the number 1. And as you progress downward in Pascal's triangle, the sum of the numbers in each row doubles compared to the sum in the row above it.

More significantly for our purposes, each number in Pascal's triangle corresponds to the combination of choosing r items out of n possibilities, typically denoted as "n choose r" or, when n equals 5 and r equals 3, "5 choose 3." The number of potential outcomes can be found at the $(n + 1)$th row and the $(r + 1)$th position in the triangle (see figure 2.1). So, given 5 and 3 for n and r, respectively, we get the number of potential outcomes by referencing the value found at the intersection of the sixth row and the fourth position in that row.

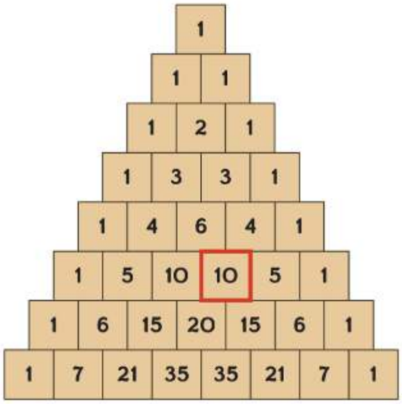

Figure 2.1 The top of Pascal's triangle. The triangle is constructed from the top down, where numbers are derived by adding the two numbers immediately above. It so happens that the sum of numbers for each row is twice the sum of numbers from the row above. The number of combinations without replacement can be found at the intersection of the (n + 1)th row and the (r + 1)th position in that row. When n equals 5 (sixth row down) and r equals 3 (fourth position from the left), the number of combinations without replacement equals 10.

In Python, we can achieve the same result in two ways: either by performing an arithmetic operation using the `factorial()` method with `n` and `r`, or by utilizing the `math.comb()` method with `n` and `r` as arguments:

```
>>> n = 5
>>> r = 3

>>> combinations = (math.factorial(n) / \
>>>                  (math.factorial(r) * math.factorial(n - r)))
>>> print(f'Number of combinations without replacement: '
>>>       f'{combinations}')
Number of combinations without replacement: 10.0

>>> combinations = math.comb(n, r)
>>> print(f'Number of combinations without replacement: '
>>>       f'{combinations}')
Number of combinations without replacement: 10
```

This brings us to combinations with replacement—maybe the most complex of all these counting rules.

COMBINATIONS WITH REPLACEMENT

A combination with replacement involves selecting items from a set where any item can potentially be selected multiple times, and the order of selection doesn't matter. Imagine purchasing a pizza topped with three out of five available ingredients. Because replacement is allowed, it's possible, for instance, to purchase extra pepperoni with mushrooms rather than pepperoni, mushrooms, and sausage. And it doesn't matter in what order these selections are made, because the pizza will be the same regardless.

The formula for calculating the number of potential combinations with replacement is actually derived from the formula for counting combinations without replacement:

$$\frac{n!}{r!\,(n-r)!}$$

When replacement is allowed, we still choose r items from a set of n distinct items, but we allow any item to be selected multiple times. To then derive the formula for combinations with replacement, we'll introduce the concept of *dummy* items.

Imagine adding $r - 1$ dummy items to the set before making selections. These are not actual items, but rather are placeholders to separate different groups of selections. By doing this, we ensure that each selection can be made independently, as the dummy items create clear divisions between groups, allowing us to account for all possible combinations, including those where some items are chosen multiple times. So, effectively, we have $n + (r - 1)$ items from which to choose; thus, $n + (r - 1))!$ becomes our numerator in lieu of $n!$ and $r!((n + (r - 1)) - r)!$ becomes our denominator in place of $r!(n - 1)!$, like so:

$$\frac{(n+(r-1))!}{r!\,((n+(r-1))-r)!}$$

or simply

$$\frac{(n+r-1)!}{r!\,(n-1)!}$$

So, if n equals 5 available ingredients and r equals 3 total selections, the number of possible combinations with replacement equals

$$\frac{(5+3-1)!}{3!\,(5-1)!}$$

or

$$\frac{7!}{3!\times 4!}$$

or

$$\frac{5,040}{6\times 24}$$

which equals 35 possible combinations.

One way of getting Python to do the heavy lifting is to again utilize the `factorial()` method with `n` and `r` as part of an arithmetic operation that mirrors the combinations with replacement formula:

```
>>> n = 5
>>> r = 3

>>> combinations = (math.factorial(n + r - 1) / \
>>>                 (math.factorial(r) * math.factorial(n - 1)))
>>> print(f'Number of combinations with replacement: '
>>>       f'{combinations}')
Number of combinations with replacement: 35.0
```

Or, even better, we can combine `n` and `r` with the `math.comb()` method:

```
>>> combinations = math.comb(n + r - 1, r)
>>> print(f'Number of combinations with replacement: '
>>>       f'{combinations}')
Number of combinations with replacement: 35
```

Now, let's shift our focus from counting rules to exploring continuous and discrete random variables.

2.3 Continuous random variables

A *random variable* is a mathematical concept that assigns a numerical value to each possible outcome of a random experiment or process. It represents uncertain quantities or events in probabilistic models, allowing for the analysis of probabilities with different outcomes. A *continuous random variable* is a variable that can take any value within a certain range or interval, including both integers (whole numbers like –8, 0, and 24) and non-integers (decimals, fractions, and mixed numbers). Unlike *discrete random variables*, which can only take on specific, countable values, continuous random variables can assume an infinite number of values within a continuous range.

Continuous random variables are characterized by their infinite precision; for instance, whereas units of time are typically measured in discrete intervals such as seconds, minutes, and hours, time itself is continuous and can therefore be measured in milliseconds, microseconds, nanoseconds, and so on. No matter how small the interval may be, there are an infinite number of possible values the variable can take on. And every possible value has a nonzero probability of occurrence.

The range of possible values for a continuous random variable is typically denoted by an interval such as $[a, b]$, where a and b represent the lower and upper bounds of the interval, respectively. For instance, the interval for the Boston Marathon is [2, 6], because the fastest runners in the world can't complete the course in less than 2 hours, and 6 hours is the cutoff time. Although every possible value between 2 and 6 hours has some probability of occurrence, it doesn't necessarily mean they have equal probabilities of occurrence. In fact, the average Boston Marathon finish time is around 4 hours, which is to say that more runners will complete the course in about 4 hours, plus or minus, but very few runners will do so in just over 2 hours or barely less than 6 hours.

Understanding continuous phenomena is essential for comprehending and predicting real-world events that occur across a spectrum of values without distinct boundaries. We'll demonstrate how to measure probabilistic behavior in many continuous phenomena, how to analyze the likelihood of different outcomes occurring, and how these probabilities accumulate across a range of values.

2.3.1 Examples

Time might be the most obvious example, but in fact, continuous random variables are used to model numerous real-world phenomena, including the following:

- Temperature measurements, whether in Celsius, Fahrenheit, or Kelvin, are continuous random variables. Temperature can take on any real value within a specified range, with infinite precision. For instance, although temperature readings of 21 °C and 72 °F are absolutely possible, so are 21.511 °C and 72.5111 °F, thereby demonstrating the continuous nature of temperature.
- Measurements of distance or length, such as the length of a road, the height of a building, or the width of a river, are continuous random variables. Distance can take on any real value within a specified range, with infinite precision. For instance, the length of a road is typically measured in kilometers or miles, but it can also be measured in fractions of a meter, millimeters, or even smaller units.
- Speed measurements, such as the speed of a vehicle or the velocity of an object, are continuous random variables. The speedometer in a car might display speeds such as 85 kilometers per hour or 55 miles per hour, and a speed gun may record the velocity of a pitched ball in a baseball game at 88 miles per hour, but speed actually takes on any real value within a specified range, with infinite precision, because speed can also be measured in fractions of kilometers or miles per hour.
- Measurements of volume or capacity, such as the volume of a container or the capacity of a reservoir, are continuous random variables. The volume of a container, for instance, may be measured in liters or gallons, but it can also be measured in fractions of a milliliter or cubic centimeter.
- Stock prices in financial markets are yet another example of continuous random variables. Any given stock may be priced at $45.00 per share, or $45.51 per share, or even $45.511 per share.

These examples are but a small subset of continuous random variables that occupy our everyday lives, whether we're aware of this simple fact or not. Although many of these and other like measurements may be presented discretely for practical purposes, it's important to acknowledge their underlying continuous nature.

2.3.2 Probability density function

The *probability density function* (PDF) is a mathematical function that quantifies the likelihood of a continuous random variable falling within a specific range of values. It is typically denoted as $f(x)$ or $p(x)$, where x represents the value of the random variable. It also provides the relative likelihood—that is, the likelihood compared to other possible values—of the random variable taking on a specific value or falling within a specific interval. So, if the interval happens to be [2, 6], meaning the lower bound is 2 and the upper bound is 6, $f(3.2)$ is the relative likelihood of observing 3.2 as the continuous random variable between 2 and 6.

The PDF is typically illustrated using a graph, where the *x* axis, or horizontal axis, represents the possible values of the random variable, and the *y* axis, or vertical axis, represents the likelihood or probability density of each value. The PDF curve is a smooth line—how smooth or not so smooth usually depends on the sample size—that exhibits different shapes depending on the specific distribution.

That is to say, the PDF is not one size fits all; the PDF formula varies depending on the specific probability distribution being modeled. Different probability distributions, such as the normal distribution versus the uniform distribution, have their own unique PDFs and therefore their own characteristics and shapes. (The uniform distribution, by the way, can represent discrete phenomena, as well; see chapter 3.)

No doubt the most common continuous probability distribution is the normal distribution, due to its prevalence among natural phenomena. In the worlds of data science and statistics, the normal distribution is predominant due to its pivotal role in a multitude of statistical techniques, including hypothesis testing and regression analysis. Thus, it only makes sense to draw a normal distribution to help further explain the PDF (see figure 2.2).

The normal distribution is a bell-shaped probability distribution characterized by its symmetric shape centered around the mean. In this particular instance—where the lower and upper bounds equal 2 and 6, respectively; the mean, or the average value, equals 4; and the standard deviation, the dispersion or spread of all values, equals 1—the distribution reveals several key properties.

For starters, the area under the PDF curve represents the probability of the continuous random variable falling within the interval [2, 6]. It illustrates the relative likelihood of observing values within the specified range.

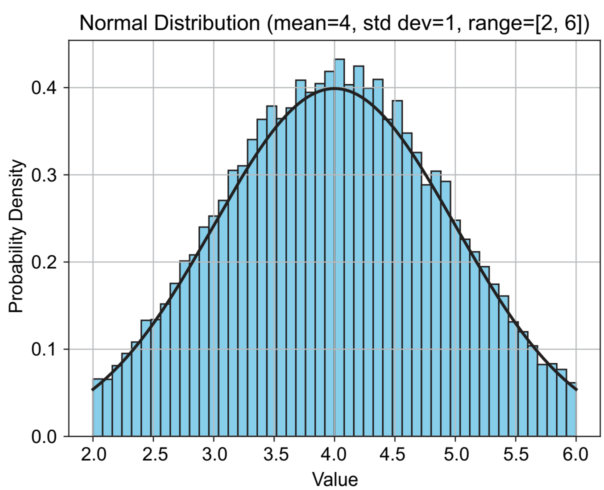

Figure 2.2 The normal distribution illustrates where the interval, or range of possible values—which runs along the *x* axis—equals [2, 6]. The probability density, which runs along the *y* axis, peaks at the mean and typically tops out at approximately 0.399. The probability density for any value can therefore be estimated merely by observation. When the value equals 3.2, the probability density appears to be equal to approximately 0.30.

Secondly, notice how the probability density is greater than 0 throughout. It peaks at the mean and approaches 0 at both tails, but it stops short of taking on a non-positive value. This guarantees that the likelihood of observing any particular value is never negative.

And finally, the integral of the PDF—that is, the total probability of all possible outcomes—is equal to the area under the curve and always sums to 1. In short, the PDF provides a method for quantifying the likelihood of different outcomes for a continuous random variable by providing insights into the distribution of probabilities across a predefined range of possible values.

Next, we'll explore the PDF's counterpart, the cumulative distribution function. It provides insights into the cumulative probability distribution of continuous random variables by integrating probabilities over a range of values.

2.3.3 Cumulative distribution function

The *cumulative distribution function* (CDF) is a function that gives the probability that a random variable takes on a value up to and including the given value x. The CDF is derived by adding up the probabilities given by the PDF. Therefore, much like the PDF, the CDF can be given as an equation and illustrated with a graph.

The CDF, typically denoted as $F(x)$ or $P(X \leq x)$, gives the probability that a continuous random variable X takes on a value less than or equal to a given point x. It represents the cumulative probability distribution of the random variable.

CDFs for continuous random variables possess many important properties. For a CDF $F(x)$ associated with the random variable X and values a and b, the following are true:

- The function $F(x)$ is nondecreasing (also called *monotonically increasing*); that is, as x increases, the probability $P(X \leq x)$ does not decrease. It will either increase or stabilize; but as a CDF, it does not and cannot ever decrease. So, if $a < b$, then $F(a) \leq F(b)$.
- The value of $F(x)$, as a probability, is always equal to some value between 0 and 1; therefore, $0 \leq F(x) \leq 1$.
- To find the probability that the random variable X takes on a value within the interval $[a, b]$, we can use the formula $P\{a < X \leq b\} = F(b) - F(a)$.
- To compute the value $F(a)$, assuming we know the PDF, we find the area under the PDF between 0 and a.

The shape of the CDF is intimately connected to that of the PDF. As the PDF changes according to the distribution being modeled, the CDF adapts accordingly, exhibiting diverse characteristics and shapes reflective of the underlying distribution. When the probability density is normally distributed, the CDF takes on an S-shaped curve (as we will see shortly).

When plotting the CDF for a normal distribution, several notable features emerge:

- The CDF starts at 0 and tops out at 1. This property actually holds true regardless of the underlying distribution.

- The CDF remains relatively flat near the tails, or when the value of the random variable is close to the lower or upper bounds of the range [2, 6]. This flatness indicates that the probabilities at the tails, as provided by the PDF, are small and contribute insignificantly to the overall CDF.
- Probabilities accumulate at a much faster rate at or around the mean. The opposite effects at the tails versus the inflection point at the mean is what gives the CDF for a normal distribution its S-shaped appearance. (It shares similarities with the sigmoid function, which will be discussed in chapter 5 during the process of fitting a logistic regression model.)
- The area under the CDF curve represents the cumulative probability. For instance, in our example, when the random variable equals 4.5, the cumulative probability—that is, the sum of probabilities given by the PDF—is approximately 70%. This means the probability of the random variable being less than 4.5 is about 70%.

All this leads to the graph in figure 2.3, where the continuous random variable is plotted along the *x* axis and the cumulative probability is plotted along the *y* axis.

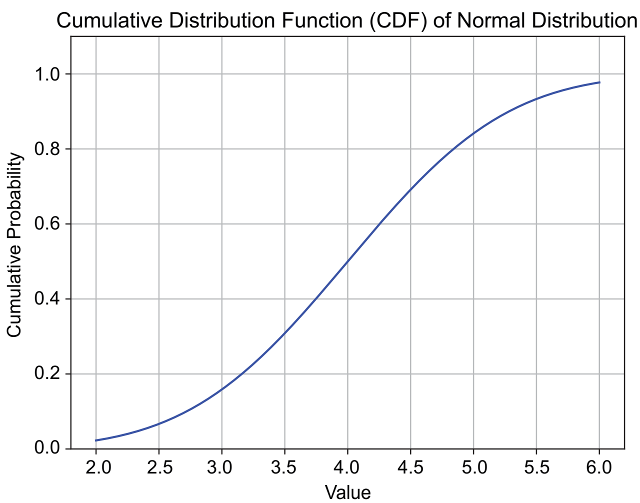

Figure 2.3 The CDF for the normal distribution when the lower and upper bounds equal 2 and 6, respectively, the mean equals 4, and the standard deviation equals 1. The shape of the CDF will vary depending on the PDF being modeled; regardless, it will always start at 0 and top out at 1.

Chapter 3 includes a more detailed discussion of the normal distribution and computing actual probabilities. For now, our focus shifts to discrete random variables.

2.4 Discrete random variables

Discrete random variables represent outcomes with distinct, countable values. For instance, when rolling a six-sided die, the whole numbers 1, 2, 3, 4, 5, and 6 represent the only possible outcomes; fractional or other "outcomes" such as 2.78 or 5.111 are impossible. This one property is what mostly differentiates discrete random variables from continuous random variables, but there are also other differences:

- Discrete random variables have a probability mass function (PMF) for calculating probabilities, whereas continuous random variables, of course, have a PDF. The PDF describes the relative likelihood of outcomes within a continuous range, whereas the PMF provides the probability of each individual outcome for a discrete random variable. Put differently, the PMF returns the actual probability of getting each specific outcome for discrete variables. The PMF for rolling a fair six-sided die assigns a probability of 1/6, or roughly 16.7%, to each possible outcome. In fact, the PMF is most commonly associated with the uniform distribution, where it assigns equal probabilities to each possible outcome within a finite range, like rolling a fair die. However, the PMF also applies to other discrete probability distributions; similar to the PDF for continuous random variables, the PMF is not one size fits all.
- The CDF for discrete random variables is a step function, increasing only at specific values, whereas the CDF for continuous random variables is instead a smooth, continuous curve. When plotted, the CDF might resemble a flight of stairs, where the height of each step represents the cumulative probability up to that outcome. For instance, where six discrete outcomes are possible and the probability is constant throughout, the cumulative probability after the first two values equals 1/6 + 1/6, or 16.7% + 16.7%, or about 33.3%.

Understanding discrete random variables is essential in probability theory and statistics, as they form the foundation for modeling discrete events, designing experiments, and making informed predictions from discrete data sets, providing valuable insights into a wide range of real-world phenomena.

2.4.1 Examples

Rolling a fair six-sided die is just one of several examples of discrete random variables. Here are a few others:

- Selecting a card from a standard deck of 52 cards. Because each card has an equal probability of being selected, 1/52, it is similar to rolling a fair die; in fact, both of these discrete random variables follow a uniform probability distribution.
- Any binomial experiment where each trial has only two possible outcomes: success or failure. For instance, this could involve counting the number of heads in a series of coin flips, where obtaining heads is deemed a success and tails a failure. Alternatively, it might entail tallying the number of defective units from a sample of items manufactured on a production line, where a defect is considered a success due to its (presumably) relative rarity compared to nondefective units.
- Scenarios where events are tallied within a set timeframe or area, like tracking the count of vehicles passing through a tollgate in an hour or the number of typographical errors found on a single page of text. In such instances, the random variable represents the frequency of (occasionally rare) events transpiring within a defined timeframe or spatial region.

- Counting the number of black marbles drawn from a bag without replacement, where the bag contains a mix of black and white marbles and a finite number of draws are allowed; or recording the number of successful applicants selected for a job from a pool of candidates, where a fixed number of candidates are selected without replacement.

These examples illustrate how discrete random variables arise in various contexts and follow different probability distributions depending on the nature of the random experiment or process being modeled.

2.4.2 Probability mass function

The PMF is to discrete random variables as the PDF is to continuous random variables. But that does not mean the PMF and PDF are otherwise alike. Formally, the PMF is denoted as $P(X = x)$, where X is the random variable and x represents its possible values. The specific formula depends on the probability distribution being modeled. Different probability distributions—binomial versus uniform versus Poisson—have their own unique PMFs, each with distinct formulas and properties. The PMF assigns an actual probability, rather than a relative likelihood, to each specific value of the random variable, which indicates the likelihood of that value occurring. Mathematically, it satisfies the following two properties:

- *Nonnegativity*—The PMF assigns nonnegative probabilities to each possible outcome, thereby ensuring that all probabilities are greater than or equal to zero.
- *Summation*—The sum of all probabilities for all possible outcomes of the random variable is equal to 1. In other words, $\Sigma P(X = x) = 1$, where the sum is taken over all possible values of x.

To illustrate, let's plot the PMF for rolling a pair of six-sided dice; see figure 2.4.

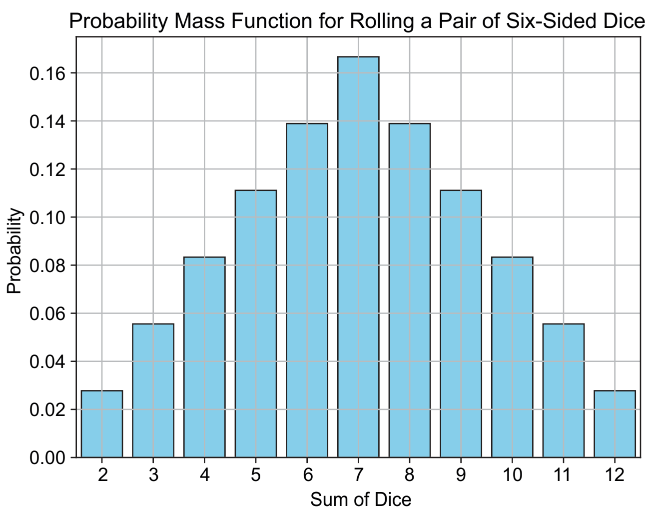

Figure 2.4 The PMF for rolling a pair of six-sided dice, where the discrete random variable and its possible values, which are equal to the whole numbers between 2 and 12, are plotted along the *x* axis, and their respective probabilities of occurrence are plotted along the *y* axis. The *y* axis represents their actual probabilities, not their relative likelihood of occurrence. All probabilities are equal to or greater than zero.

The discrete random variable and its possible values—that is, every possible outcome from rolling a pair of six-sided dice, which are and must be in the form of integers or whole numbers—are plotted along the *x* axis. Their respective probabilities—not their relative likelihood of occurrence—are plotted along the *y* axis. Every individual probability is, and must be, equal to or greater than zero.

For instance, out of 36 possible outcomes, there is only one way to get a pair of six-sided dice to total 2, so

$$P(X = 2) = \frac{1}{36}$$

or

$$P(X = 2) = 0.0278$$

But there are two ways to get a pair of six-sided dice to total 3: a 1 from one die and a 2 from the other, or vice versa. So

$$P(X = 3) = \frac{2}{36}$$

or

$$P(X = 3) = 0.0556$$

And there are exactly three ways of getting the same dice to total 4, so

$$P(X = 4) = \frac{3}{36}$$

or

$$P(X = 4) = 0.0833$$

To reiterate, these values represent actual probabilities. The PDF for continuous random variables returns relative likelihoods, but the PMF for discrete random variables returns actual probabilities of occurrence for each possible value.

2.4.3 Cumulative distribution function

The CDF of a discrete random variable X, typically denoted as $F(x)$, is defined as the probability that X is less than or equal to a certain value x. It is more formally expressed as $F(x) = P(X \leq x)$.

It assumes the following properties:

- *Nondecreasing*—The CDF is nondecreasing, meaning that as the value of x increases, the cumulative probability either increases or at least remains constant. It will not and cannot ever decrease.
- *Bounded*—The CDF is bounded between 0 and 1, inclusive. This reflects the fact that probabilities always range from 0 to 1.
- *Right-continuous*—The CDF is typically right-continuous, meaning that the cumulative probability jumps at each value of x, reflecting the total probability up to and including that point, with no jumps or increases between values of x.

The CDF is calculated by simply summing the probabilities of all possible values less than or equal to a specific value of x. So, $F(x) = \mathrm{P}(X \leq x) = \Sigma \mathrm{P}(X = k)$, where the sum is taken over all possible values of k that are less than or equal to x.

Consider the cumulative probabilities of rolling a pair of six-sided dice. The probability of getting 4 or less is equal to the probability of getting 2 plus the probability of getting 3 plus the probability of getting 4:

$$\mathrm{P}(X = 2) = .0278$$

$$+$$

$$\mathrm{P}(X = 3) = .0556$$

$$+$$

$$\mathrm{P}(X = 4) = .0833$$

$$=$$

$$F(4) = 0.167$$

Thus, the probability of getting 4 or less from rolling a pair of six-sided dice equals 16.7%.

As mentioned previously, the CDF is typically represented in graphical format using a step function (see figure 2.5). The plot consists of horizontal line segments connecting consecutive values for x, with each segment representing the cumulative probability up to that point.

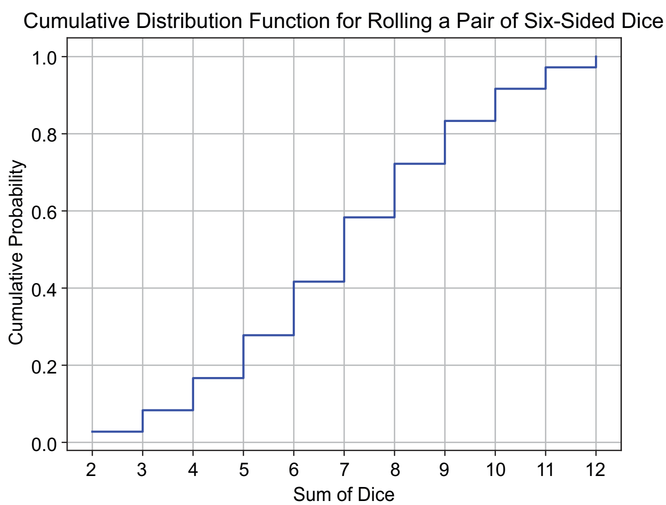

Figure 2.5 The CDF for rolling a pair of six-sided dice. The shape of the distribution resembles a two-dimensional flight of stairs in which the horizontal line segments represent consecutive values for *x* and the vertical lines represent their respective probabilities.

To further explain, the probability of getting 2 equals 0.0278, which is exactly where the horizontal line segment connecting 2 and 3 is drawn. The probability of getting 3 equals 0.0556; rather than it being drawn from 0, it is instead drawn from where the horizontal line segment ends. Therefore, the point at which the vertical line ends is the cumulative probability of getting 3 or less. This process of adding each probability to the previous cumulative value is repeated for each successive outcome until we reach the last possible outcome, and the cumulative probability at that point equals 1.

As we conclude this chapter and prepare for the next, let's quickly reflect on all the content we packed into these pages and briefly summarize what we've learned. We learned how to compute a theoretical probability based on mathematical principles and assumptions and then how to compute an empirical probability derived from observed data and real-world experiments. We discovered the fundamental principles of the multiplication and addition rules of counting, equipping us with the knowledge of their precise applications and significance across various probability scenarios. We explored the intricacies of combinations and permutations, including how to set them apart, and we are now familiar with the concepts of replacement and without replacement, and specifically how one leads to a greater number of combinations or permutations than the other. And we examined the characteristics of both continuous and discrete random variables, shedding light on their unique properties and applications.

This sets us up well for chapter 3. The next chapter picks up right where we are about to leave off: with a discussion of various probability distributions.

Summary

- Theoretical probabilities are equal to the number of successes divided by the total number of possible outcomes.
- Empirical probabilities are derived from trials or real-world observations. Such probabilities are equal to the number of successes observed (however a "success" might be defined) divided by the number of observations made.
- The multiplication rule states that the probability of two or more events occurring is equal to the product of their individual probabilities. This rule applies to independent events that occur simultaneously or sequentially.
- The addition rule states that the probability of two or more mutually exclusive events occurring is equal to the sum of their individual probabilities. This rule applies when the events cannot occur simultaneously.
- Combinations and permutations are not the same. A combination refers to the selection of items where the order does not matter, whereas a permutation involves arrangements where the order means everything. Combinations apply to scenarios like choosing a group of students to sit on a committee, and permutations apply to situations like arranging a sequence of numbers or letters.
- A permutation with replacement is a method of arranging items from a set where each item can be chosen multiple times for each position in the

arrangement. This allows for the repetition of items in the arrangement, resulting in a larger number of possible permutations compared to permutations without replacement.

- A permutation without replacement is a method of arranging items from a set where each item can be chosen only once for each position in the arrangement. This ensures that each item appears no more than once in the final arrangement, leading to fewer permutations compared to permutations with replacement.
- A combination without replacement is a selection of items from a set where each item can be selected no more than once, and the order of selection does not matter. This ensures that each item is selected maybe once for the combination, leading to a unique subset of items from the original set.
- A combination with replacement is a selection of items from a set where each item can be chosen multiple times, and the order of selection does not matter. This allows for the repetition of items in the combination, resulting in a larger number of possible combinations compared to combinations without replacement.
- A continuous random variable is a type of random variable that can take on any value within a certain range, often representing measurements or quantities that can be infinitely subdivided.
- A discrete random variable is a type of random variable that can take on only a countable number of distinct values, often representing outcomes of experiments or events that result in a finite or countably infinite set of possible outcomes. Unlike continuous random variables, which can assume any value within a range, discrete random variables have specific and separate values with no intermediate possibilities.