## Contents

### 2 Probability

- 2.1 Introduction ................................................................. 1
- 2.2 Sample spaces and sets .................................................. 6
- 2.3 Axioms of probability ................................................... 9
- 2.4 Discrete sample spaces .................................................. 12
- 2.5 Tools for counting outcomes (sample points) ...................... 15
  - 2.5.1 Basic counting rule .................................................. 15
  - 2.5.2 Permutations ......................................................... 17
  - 2.5.3 Multinomial coefficients .......................................... 19
  - 2.5.4 Combinations ........................................................ 22
- 2.6 Conditional probability and independence ........................ 25
- 2.7 Law of Total Probability and Bayes’ Rule ......................... 31

### 3 Discrete Random Variables and their Probability Distributions 35

- 3.1 Introduction ................................................................. 35
- 3.2 Mathematical expectation .............................................. 42
  - 3.2.1 Expected value ....................................................... 42
  - 3.2.2 Functions of random variables ................................. 46
  - 3.2.3 Variance .............................................................. 47
- 3.3 Moment-generating functions ....................................... 50
- 3.4 Binomial distribution .................................................... 53
- 3.5 Geometric distribution .................................................. 58
- 3.6 Negative binomial distribution ...................................... 61
- 3.7 Hypergeometric distribution ......................................... 64
- 3.8 Poisson distribution ...................................................... 68
2 Probability

2.1 Introduction

Terminology: Probability is a measure of one’s belief in the occurrence of a future event. Probability is called “the mathematics of uncertainty.”

Examples: Here are some events we might want to assign a probability to (i.e., to quantify the likelihood of occurrence):

- tomorrow’s temperature exceeding 80 degrees
- getting a flat tire on my way home today
- a new policy holder making a claim in the next year
- you being diagnosed with prostate/cervical cancer in the next 20 years
- a new patient developing an addiction to opioids
- President Trump winning re-election in 2020.

Approaches: How do we assign probabilities to events like these and other events?

1. Subjective approach
   - based on prior experience, subject-matter knowledge, feeling, etc.
   - may not be scientific

2. Relative frequency approach
   - requires the ability to observe the occurrence of an event (and its non-occurrence) repeatedly under identical conditions
   - can be carried out using simulation (see Examples 2.1, 2.2, and 2.3)

3. Axiomatic/Model-based approach
   - grounded in set theory/mathematics
   - we will take this approach

Example 2.1. We illustrate how the relative frequency approach works using simulation. Suppose we flip a coin and observe the outcome; the sample space is

\[ S = \{H, T\}. \]

Let \( A = \{T\} \), the event that a “tail” is observed. How might we assign probability to the event \( A \)? Suppose we flip the same coin over and over again and record the fraction of
times $A$ occurs. This fraction is called the **relative frequency**. Specifically, if we flip the coin $n$ times and let $n(A)$ denote the number of times $A$ occurs, then the relative frequency approach to probability says

$$P(A) \approx \frac{n(A)}{n}.$$ 

The symbol $P(A)$ is shorthand for “the probability that $A$ occurs.”

**Illustration:** I used R to simulate this experiment $n = 10000$ times while assuming the coin is fair; i.e., flipping the same fair coin 10000 times. Figure 2.1 plots the relative frequencies over the 10000 flips. The number of tails observed in this simulation was 5040.

```r
> sum(flip)
[1] 5040
```

Therefore, we would assign

$$P(A) = 0.5040$$

on the basis of this simulation. If we repeated the simulation, we would get different answers most likely. In fact, I did this 5 more times and got 5023, 5033, 5016, 5061, and 4976.
Remark: In general, the relative frequency approach to probability says that \( n(A)/n \) will “stabilize” around \( P(A) \) as \( n \) increases. Mathematically,

\[
\lim_{n \to \infty} \frac{n(A)}{n} = P(A).
\]

Interesting: John Edmund Kerrich (a British mathematician) performed a similar experiment. He flipped an actual fair coin \( n = 10000 \) times while in an internment camp in Nazi-occupied Denmark in the 1940’s (he did not have R!). He observed 5067 heads out of 10000 flips, offering empirical evidence of why the relative frequency approach “works” (as we have just done). □

Example 2.2. The matching problem. Suppose \( M \) men are at a party, and each man is wearing a hat. Each man throws his hat into the center of the room. Each man then selects a hat at random. What is the probability at least one man selects his own hat; i.e., there is at least one “match”? Define

\[ A = \{ \text{at least one man selects his own hat} \}. \]

Let’s use simulation to estimate \( P(A) \) like we did in Example 2.1.

Illustration: I used R to perform the “hat matching” experiment \( n = 10000 \) times while assuming the party consisted of \( M = 10 \) men. The event \( A \) occurred in 6364 of the simulated parties:

\[
\text{> sum(event)}
\]

\[
[1] 6364
\]

Therefore, we would assign

\[ P(A) = 0.6364 \]

on the basis of this simulation. The plot of relative frequencies is shown in Figure 2.2 (next page).

Curiosity: What happens if we grow the size of the party? I performed the same simulation with \( M = 100, M = 1000, \) and \( M = 10000 \) men and obtained the following results:

<table>
<thead>
<tr>
<th>( M )</th>
<th>( n(A) )</th>
<th>( n )</th>
<th>( P(A) = n(A)/n )</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>6364</td>
<td>10000</td>
<td>0.6364</td>
</tr>
<tr>
<td>100</td>
<td>6342</td>
<td>10000</td>
<td>0.6342</td>
</tr>
<tr>
<td>1000</td>
<td>6300</td>
<td>10000</td>
<td>0.6300</td>
</tr>
<tr>
<td>10000</td>
<td>6351</td>
<td>10000</td>
<td>0.6351</td>
</tr>
</tbody>
</table>

Interesting: In general, for a party with \( M \) men, the probability of at least one match is

\[
P(A) = 1 - \sum_{k=0}^{M} \frac{(-1)^k}{k!}.
\]
Letting the party grow large without bound is equivalent to letting $M \to \infty$. From calculus,

$$
\lim_{M \to \infty} \left[ 1 - \sum_{k=0}^{M} \frac{(-1)^k}{k!} \right] = 1 - \sum_{k=0}^{\infty} \frac{(-1)^k}{k!} = 1 - e^{-1} \approx 0.6321.
$$

Remark: This is an example where intuition usually leads one astray. Some students would reason that as the number of men $M$ increases, the chance of an individual match ($1/M$) decreases to 0 so $P(A)$ will also approach 0. Other students would reason that because $M$ is large, “hat matching” overall becomes easier so $P(A)$ should approach 1. Neither argument is correct and, in fact, the correct answer lies somewhere in the middle.

Example 2.3. The birthday problem. In a class of $M$ students, what is the probability there will be at least one shared birthday? Define

$$
A = \{ \text{at least one shared birthday} \}.
$$

Let’s use simulation to estimate $P(A)$. To make this example concrete, assume that there are 365 days in a year and that there are no siblings (e.g., twins, triplets, etc.) in the class. On July 1, 2018, there were $M = 50$ students enrolled in this class, so we will use this.
Illustration: I used R to perform the “shared birthday” experiment $n = 10000$ times while assuming the class consists of $M = 50$ students. The event $A$ occurred in 9697 of the simulated parties:

> sum(event)
[1] 9697

Therefore, we would assign

$$P(A) = 0.9697$$

on the basis of this simulation. The plot of relative frequencies is shown in Figure 2.3.

Interesting: In general, for a class with $M$ students, the correct answer is

$$P(A) = 1 - \frac{M! (365)^M}{365^M}.$$  

When $M = 50$, this probability is 0.9704 (to 4 dp), so our simulation was accurate. Interestingly, $M$ need only be 23 for $P(A)$ to exceed 1/2. Intuition might suggest that you would need many more students than this. □
2.2 Sample spaces and sets

**Terminology:** A random experiment is an experiment that produces outcomes which are not predictable with certainty in advance. The sample space $S$ for a random experiment is the set of all possible outcomes.

**Example 2.4.** Consider the following random experiments and their associated sample spaces. Let $\omega$ denote a generic outcome.

(a) Observe the high temperature for today:

$$S = \{ \omega : -\infty < \omega < \infty \} = \mathbb{R}$$

(b) Record the number of planes landing at CAE:

$$S = \{ \omega : \omega = 0, 1, 2, \ldots \} = \mathbb{Z}^+$$

(c) Toss a coin three times:

$$S = \{(\text{HHH}), (\text{HHT}), (\text{HTH}), (\text{THH}), (\text{THT}), (\text{TTH}), (\text{TTT})\}$$

(d) Measure the length of a female subject’s largest uterine fibroid:

$$S = \{ \omega : \omega \geq 0 \} = \mathbb{R}^+$$

**Definitions:** We say that a set (e.g., $A$, $B$, $S$, etc.) is countable if its elements can be put into a 1:1 correspondence with the set of natural numbers

$$\mathbb{N} = \{1, 2, 3, \ldots \}. $$

If a set is not countable, we say it is uncountable. In Example 2.4,

(a) $S = \mathbb{R}$ is uncountable

(b) $S = \mathbb{Z}^+$ is countable (i.e., countably infinite); $|S| = +\infty$

(c) $S = \{(\text{HHH}), (\text{HHT}), \ldots, (\text{TTT})\}$ is countable (i.e., countably finite); $|S| = 8$

(d) $S = \mathbb{R}^+$ is uncountable

**Note:** Any finite set is countable. By “finite,” we mean that $|A| < \infty$, that is, “the process of counting the elements in $A$ comes to an end.” An infinite set $A$ can be countable or uncountable. By “infinite,” we mean that $|A| = +\infty$. For example,

- $\mathbb{N} = \{1, 2, 3, \ldots \}$ is countably infinite
- $A = \{ \omega : 0 < \omega < 1 \}$ is uncountable.
Definitions: Suppose that $S$ is a sample space for a random experiment. An event $A$ is a subset of $S$, that is, $A \subseteq S$. Suppose the experiment produces the outcome $\omega$.

- If $\omega \in A$, we say that “$A$ occurs”
- If $\omega \notin A$, we say that “$A$ does not occur.”

The set $A$ is a subset of $B$ if $$\omega \in A \implies \omega \in B.$$ This is written $A \subset B$ or $A \subseteq B$. In a random experiment, if the event $A$ occurs, then so does $B$. The converse is not necessarily true.

Two sets $A$ and $B$ are equal if each set is a subset of the other, that is, $$A = B \iff A \subseteq B \text{ and } B \subseteq A.$$ In probability, set equality is important. If two events $A$ and $B$ are the same (i.e., $A = B$), then they have the same probability.

The null set, denoted by $\emptyset$, is the set that contains no outcomes. Intuitively, it makes sense to assign zero probability to this “event.”

Set Operations: Suppose $A$ and $B$ are subsets of $S$.

- Union: $A \cup B = \{\omega \in S : \omega \in A \text{ or } \omega \in B\}$. This is the set of all outcomes in $A$ or in $B$ (or in both).
- Intersection: $A \cap B = \{\omega \in S : \omega \in A \text{ and } \omega \in B\}$. This is the set of all outcomes in $A$ and $B$.
- Complementation: $\overline{A} = \{\omega \in S : \omega \notin A\}$. This is the set of all outcomes not in $A$.

Example 2.5. A medical professional observes adult male patients entering an emergency room. She classifies each patient according to his blood type ($\text{AB}^+, \text{AB}^-, \text{A}^+, \text{A}^-, \text{B}^+, \text{B}^-, \text{O}^+, \text{and} \text{O}^-$) and whether his systolic blood pressure (SBP) is low (L), normal (N), or high (H). Consider the observation of the next male patient as a random experiment.

The sample space is

$$S = \{(\text{AB}^+, \text{L}), (\text{AB}^-, \text{L}), (\text{A}^+, \text{L}), (\text{A}^-, \text{L}), (\text{B}^+, \text{L}), (\text{B}^-, \text{L}), (\text{O}^+, \text{L}), (\text{O}^-, \text{L}), (\text{AB}^+, \text{N}), (\text{AB}^-, \text{N}), (\text{A}^+, \text{N}), (\text{A}^-, \text{N}), (\text{B}^+, \text{N}), (\text{B}^-, \text{N}), (\text{O}^+, \text{N}), (\text{O}^-, \text{N}), (\text{AB}^+, \text{H}), (\text{AB}^-, \text{H}), (\text{A}^+, \text{H}), (\text{A}^-, \text{H}), (\text{B}^+, \text{H}), (\text{B}^-, \text{H}), (\text{O}^+, \text{H}), (\text{O}^-, \text{H})\}.$$ Note that this sample space contains $|S| = 24$ outcomes.
This example illustrates many concepts we will discuss in due course:

- There are 8 different blood types. There are 3 different categorizations of SBP. There are \(8 \times 3 = 24\) possible outcomes in the sample space, which is formed by combining the two factors. The authors call this “the \(mn\) rule.”

- Because \(S\) is countable, the authors call this a **discrete sample space**.

- Are these 24 outcomes equally likely? Probably not. O\(^+\) is by far the most common blood type among American males (about 38 percent). On the other hand, AB\(^-\) is rare (only about 1 percent). Similarly, most American males have either normal or high SBP; much fewer have low SBP.

- Even though we have listed all possible outcomes in \(S\), we have not specified probabilities associated with the outcomes. We cannot assign probability to events like 
  
  \[
  A = \{\text{blood type with a } + \text{ rhesus status}\} \\
  B = \{\text{high SBP}\}
  
  \]

  without having this information.

**Exercise:** List the outcomes in \(A \cup B\), \(A \cap B\), and \(\overline{A}\).

\[
A \cup B = \{\text{outcomes with a } + \text{ rhesus status or high SBP}\} \\
= \{(AB^+, L), (A^+, L), (B^+, L), (O^+, L), (AB^+, N), (A^+, N), (B^+, N), (O^+, N), (AB^+, H), (AB^-, H), (A^+, H), (A^-, H), (B^+, H), (B^-, H), (O^+, H), (O^-, H)\}

A \cap B = \{\text{outcomes with a } + \text{ rhesus status and high SBP}\} \\
= \{(AB^+, H), (A^+, H), (B^+, H), (O^+, H)\}

\overline{A} = \{\text{outcomes with a } - \text{ rhesus status}\} \\
= \{(AB^-, L), (A^-, L), (B^-, L), (O^-, L), (AB^-, M), (A^-, M), (B^-, M), (O^-, M), (AB^-, H), (A^-, H), (B^-, H), (O^-, H)\}

**Exercise:** List the outcomes in \(\overline{A} \cup B\), \(A \cap \overline{B}\), and \(\overline{A} \cap \overline{B}\). \(\square\)

Here are two last set theory results that will prove to be useful.

**Distributive Laws:**

\[
A \cap (B \cup C) = (A \cap B) \cup (A \cap C) \\
A \cup (B \cap C) = (A \cup B) \cap (A \cup C)
\]

**DeMorgan’s Laws:**

\[
\overline{A} \cup B = \overline{A} \cap B \\
\overline{A} \cap B = \overline{A} \cup B
\]
2.3 Axioms of probability

Terminology: We say that two events $A$ and $B$ are **mutually exclusive** or **disjoint** if

\[ A \cap B = \emptyset, \]

that is, $A$ and $B$ have no outcomes in common. Mutually exclusive events cannot occur simultaneously. For example, clearly $A$ and $\overline{A}$ are mutually exclusive. If $A$ occurs, then $\overline{A}$ cannot occur and vice versa.

**Kolmogorov’s Axioms:** Suppose $S$ is a sample space and let $\mathcal{B}$ denote the collection of all possible events. Let $P$ be a set function; i.e.,

\[ P : \mathcal{B} \rightarrow [0, 1], \]

that satisfies the following axioms:

1. $P(A) \geq 0$, for all $A \in \mathcal{B}$
2. $P(S) = 1$
3. If $A_1, A_2, ..., \in \mathcal{B}$ are pairwise mutually exclusive; i.e., $A_i \cap A_j = \emptyset \forall i \neq j$, then

\[ P \left( \bigcup_{i=1}^{\infty} A_i \right) = \sum_{i=1}^{\infty} P(A_i). \]

We call $P$ a **probability set function** (or probability measure).

**Remark:** Mathematically, we can think of $P$ as a function whose domain is sets (events) and whose range is $[0, 1]$. Therefore, probabilities are numbers between 0 and 1 (inclusive). In a more advanced course, one would describe the collection of events $\mathcal{B}$ much more carefully to avoid certain peculiar mathematical contradictions; we will not.

**Consequences:** From these 3 axioms, we can develop numerous rules which help us assign probability to events.

1. **Complement rule:** $P(A) = 1 - P(\overline{A})$, for any event $A$.
   
   **Proof.** We can write $S = A \cup \overline{A}$. Because $A$ and $\overline{A}$ are mutually exclusive, Axiom 3 says
   
   \[ P(S) = P(A \cup \overline{A}) = P(A) + P(\overline{A}). \]

   However, $P(S) = 1$ by Axiom 2. Therefore,

   \[ P(A) = 1 - P(\overline{A}). \]

   **Importance:** In many problems, it is often much easier to calculate the probability that $A$ does not occur. If you can do this, then the complement rule gives $P(A)$ easily.
2. **Null set rule:** $P(\emptyset) = 0.$

   **Proof.** This follows immediately from the complement rule; take $A = \emptyset$ and $\overline{A} = S.$ \(\square\)

3. **Upper bound rule:** $P(A) \leq 1.$

   **Proof.** In the proof of the complement rule, we wrote
   
   $$P(S) = P(A \cup \overline{A}) = P(A) + P(\overline{A}).$$
   
   However, $P(\overline{A}) \geq 0$ by Axiom 1 and $P(S) = 1$ by Axiom 2. \(\square\)

4. **Monotonicity rule:** Suppose $A$ and $B$ are events such that $A \subset B,$ that is, $A$ is a subset of $B.$ Then
   
   $$P(A) \leq P(B).$$
   
   This result makes sense intuitively. If $A$ occurs, then $B$ must occur. However, the reverse is not true, so $P(B)$ must be larger (or at least not smaller).

   **Proof.** Because $A \subset B,$ we can write
   
   $$B = A \cup (B \cap \overline{A});$$
   
   i.e., $B \cap \overline{A}$ captures all outcomes in $B$ and not in $A.$ Clearly, $A$ and $(B \cap \overline{A})$ are mutually exclusive. Thus, from Axiom 3, we have
   
   $$P(B) = P(A) + P(B \cap \overline{A}).$$
   
   However, from Axiom 1, $P(B \cap \overline{A}) \geq 0,$ so $P(B) \geq P(A).$ \(\square\)

5. **Additive rule:** Suppose $A$ and $B$ are two events.

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

   **Remark:** We know if $A$ and $B$ are mutually exclusive, then
   
   $$P(A \cup B) = P(A) + P(B).$$
   
   This is what Axiom 3 guarantees. So the additive rule is more general; i.e., $A$ and $B$ need not be mutually exclusive.

   **Proof.** Write $A \cup B = A \cup (\overline{A} \cap B).$ Because $A$ and $(\overline{A} \cap B)$ are mutually exclusive,
   
   $$P(A \cup B) = P(A) + P(\overline{A} \cap B)$$
   
   by Axiom 3. Now, write $B = (A \cap B) \cup (\overline{A} \cap B).$ Clearly, $(A \cap B)$ and $(\overline{A} \cap B)$ are mutually exclusive too. From Axiom 3 again,
   
   $$P(B) = P(A \cap B) + P(\overline{A} \cap B).$$
   
   Combining the expressions for $P(\overline{A} \cap B)$ in both equations above gives the result. \(\square\)
Example 2.6. A smoke detector system uses two interlinked units. If smoke is present, the probability the first unit will detect it is 0.95 and the probability the second unit will detect it is 0.90. The probability smoke will be detected by both units is 0.88.

(a) If smoke is present, find the probability that the smoke will be detected by either unit or both.
(b) Find the probability the smoke will go undetected.

Solutions. Define the events

\[ A = \{ \text{first unit detects smoke} \} \]
\[ B = \{ \text{second unit detects smoke} \}. \]

We are given \( P(A) = 0.95 \), \( P(B) = 0.90 \), and \( P(A \cap B) = 0.88 \).

(a) The probability the system will detect smoke (when present) is

\[
P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.95 + 0.90 - 0.88 = 0.97.
\]

(b) Smoke will go undetected when \( A \cap B \) occurs. By DeMorgan’s Law,

\[
P(A \cap B) = P(A \cup B) = 1 - P(A \cup B) = 0.03. \quad \square
\]

Remark: The additive rule can be generalized for any sequence of events \( A_1, A_2, ..., A_n \); i.e.,

\[
P\left( \bigcup_{i=1}^{n} A_i \right) = \sum_{i=1}^{n} P(A_i) - \sum_{i_1 < i_2} P(A_{i_1} \cap A_{i_2}) + \sum_{i_1 < i_2 < i_3} P(A_{i_1} \cap A_{i_2} \cap A_{i_3}) - \cdots + (-1)^{n+1} P\left( \bigcap_{i=1}^{n} A_i \right).
\]

For example, if \( n = 3 \), then

\[
P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3) - P(A_1 \cap A_2) - P(A_1 \cap A_3) - P(A_2 \cap A_3) + P(A_1 \cap A_2 \cap A_3).
\]

Example 2.7. Prove Bonferroni’s Inequality; i.e.,

\[
P(A \cap B) \geq 1 - P(\overline{A}) - P(\overline{B}).
\]

Proof. From the additive rule and complement rule, we know

\[
P(A \cap B) = P(A) + P(B) - P(A \cup B) = [1 - P(\overline{A})] + [1 - P(\overline{B})] - P(A \cup B) = 1 - P(\overline{A}) - P(\overline{B}) + [1 - P(A \cup B)].
\]

However, \( 1 - P(A \cup B) = P(\overline{A} \cup \overline{B}) \) is itself a probability and hence \( 1 - P(A \cup B) \geq 0 \) by Axiom 1. Thus, we are done. \( \square \)
**Generalization:** Bonferroni’s Inequality can be generalized for any sequence of events \(A_1, A_2, ..., A_n\); i.e., \[ P\left(\bigcap_{i=1}^{n} A_i\right) \geq 1 - \sum_{i=1}^{n} P(A_i). \]

**Application:** Bonferroni’s Inequality is useful in statistics when multiple confidence intervals are being written. In this context, the event \(A_i = \{i\text{th interval includes its target parameter}\}\) and the event \(\bigcap_{i=1}^{n} A_i = \{\text{each confidence interval includes its target parameter}\}\).

For example, if \(n = 5\) and \(P(A_i) = 0.95\) (confidence level), then the probability all 5 intervals will include their target parameter is \[ P\left(\bigcap_{i=1}^{5} A_i\right) \geq 1 - 5(0.05) = 0.75. \] This probability, which corresponds to the family of 5 intervals taken together, can be much lower than each interval’s confidence level of 0.95. Furthermore, the fact that this “familywise confidence level” can be so low is concerning.

### 2.4 Discrete sample spaces

**Terminology:** Suppose \(S\) is a sample space for a random experiment. If \(S\) contains a finite or countable number of outcomes, we call \(S\) a **discrete sample space**. Recall:

- **Finite:** \(|S| < \infty\); i.e., the number of outcomes in \(S\) is finite
- **Countable:** the number of outcomes may be infinite; i.e., \(|S| = +\infty\), but the outcomes can be put into a 1:1 correspondence with \(\mathbb{N} = \{1, 2, 3, \ldots\}\).

**Example 2.8.** An American style roulette wheel contains 38 numbered compartments or “pockets.” The pockets are either red, black, or green. The numbers 1 through 36 are evenly split between red and black, while 0 and 00 are green pockets. Conceptualize the next spin of the wheel as a random experiment with sample space \(S = \{1, 2, 3, 4, ..., 34, 35, 36, 0, 00\}\).

Note that this is a discrete sample space with \(|S| = 38\) outcomes (sample points).
Consider the following events (i.e., subsets of $S$):

\[ A_1 = \{13\} \]
\[ A_2 = \{\text{“red”}\} = \{1, 3, 5, 7, 9, 12, 14, 16, 18, 19, 21, 23, 25, 27, 30, 32, 34, 36\} \]
\[ A_3 = \{\text{“green”}\} = \{0, 00\}. \]

**Terminology:** A **simple event** is an event that consists of exactly one outcome (sample point).

- In Example 2.8, the event $A_1 = \{13\}$ is a simple event.

A **compound event** is an event that contains more than one outcome (sample point). Therefore, any compound event can be written as the (countable) union of simple events. In Example 2.8, the event $A_3 = \{\text{“green”}\} = \{0, 00\}$ can be written as

\[ A_3 = \{0\} \cup \{00\}, \]

the union of 2 simple events. The event $A_2 = \{\text{“red”}\}$ can be written as the union of 18 simple events.

**Important:** In a discrete sample space, calculating the probability of a compound event $A$ is done by adding up the probabilities associated with each sample point in it. That is,

\[ P(A) = \sum_{i: E_i \subset A} P(E_i), \]

where $E_1, E_2, \ldots, E_n$ denote the simple events whose union makes up $A$. This strategy to calculate $P(A)$ follows from Axiom 3. If the compound event $A$ can be expressed as $A = E_1 \cup E_2 \cup \cdots \cup E_n$ (for a finite number of simple events), then

\[ P(A) = P(E_1 \cup E_2 \cup \cdots \cup E_n) = P(E_1) + P(E_2) + \cdots + P(E_n). \]

**Example 2.9.** Consider the random experiment of observing the number of children born during the next live birth in the United States. A sample space for this experiment is

\[ S = \{1, 2, 3, 4, 5+\}. \]

Note that this is a discrete sample space with $|S| = 5$ outcomes (sample points).

Let $E_1, E_2, E_3, E_4, E_5$ denote the five simple events that make up $S$. The CDC reports the following probabilities during 2015 (among all 3,978,497 live births):

\[ P(E_1) = 0.965399 \]
\[ P(E_2) = 0.033489 \]
\[ P(E_3) = 0.001047 \]
\[ P(E_4) = 0.000059 \]
\[ P(E_5) = 0.000006 \]

It is easy to see that $P(E_1) + P(E_2) + P(E_3) + P(E_4) + P(E_5) = 1$, as it should (Axiom 2).
Q: Under this model, what is the probability of a multiple birth? Note that the event \( A = \{ \text{multiple birth} \} \) can be written as the union of the 4 simple events:

\[ A = E_2 \cup E_3 \cup E_4 \cup E_5. \]

Therefore,

\[
P(A) = P(E_2) + P(E_3) + P(E_4) + P(E_5) = 0.033489 + 0.001047 + 0.000059 + 0.000006 = 0.034601.
\]

Of course, using the complement rule gives you the same answer:

\[
P(A) = 1 - P(E_1) = 1 - 0.965399 = 0.034601. \]

\[ \square \]

**Example 2.10.** Two jurors are needed to serve as “alternates” in an attempted murder trial. These two jurors will be selected from 5 potential jurors, three men and two women. Envision the selection of these two jurors as a random experiment with sample space

\[ S = \{(M_1, M_2), (M_1, M_3), (M_1, W_1), (M_1, W_2), (M_2, M_3), (M_2, W_1), (M_2, W_2), (M_3, W_1), (M_3, W_2), (W_1, W_2)\}. \]

Note that this is a discrete sample space with \( |S| = 10 \) outcomes (sample points).

Q: What is the probability at least one woman is selected as an alternate juror? We do not have enough information to answer this question because we don’t know the probabilities associated with the 10 sample points. However, certainly we can list out the sample points in this event:

\[
A = \{ \text{at least one woman selected} \} = \{(M_1, W_1), (M_1, W_2), (M_2, W_1), (M_2, W_2), (M_3, W_1), (M_3, W_2), (W_1, W_2)\}.
\]

**Note:** If we assume the outcomes in \( S \) are equally likely; i.e., each with probability

\[
\frac{1}{|S|} = \frac{1}{10},
\]

then we can compute \( P(A) \). It is simply

\[
P(A) = \frac{\text{number of outcomes in } A}{\text{number of outcomes in } S} = \frac{7}{10}.
\]

However, it is important to understand that this simple rule for assigning probabilities is only valid when the outcomes in \( S \) are equally likely. If the outcomes in \( S \) are not equally likely, then this probability assignment is not correct.
Equi-probability model: Suppose the discrete sample space $S$ contains $N = |S| < \infty$ outcomes (sample points), each of which is equally likely. If the event $A$ contains $n_a$ outcomes (sample points), then

\[ P(A) = \frac{n_a}{N}. \]

Proof. Write $S = E_1 \cup E_2 \cup \cdots \cup E_N$, where $E_i$ denotes the $i$th simple event, for $i = 1, 2, \ldots, N$. Then,

\[ 1 = P(S) = P(E_1 \cup E_2 \cup \cdots \cup E_N) = \sum_{i=1}^{N} P(E_i), \]

by Axioms 2 and 3. Because $P(E_1) = P(E_2) = \cdots = P(E_N)$,

\[ 1 = \sum_{i=1}^{N} P(E_i) \implies P(E_i) = \frac{1}{N}, \quad i = 1, 2, \ldots, N. \]

Without loss of generality, take $A = E_1 \cup E_2 \cup \cdots \cup E_{n_a}$. Then,

\[ P(A) = P(E_1 \cup E_2 \cup \cdots \cup E_{n_a}) = \sum_{i=1}^{n_a} P(E_i) = \sum_{i=1}^{n_a} \frac{1}{N} = \frac{n_a}{N}. \]

**Implication:** Suppose $S$ is a discrete sample space with a finite number of outcomes; i.e., $N < \infty$. If each outcome is equally likely, then finding $P(A)$ reduces to two “counting problems:” one to find $N$ and one to find $n_a$.

- In simple experiments, like Example 2.10, we can simply list out all outcomes in $S$ and $A$ and count to find $N$ and $n_a$ quickly.

- In more complicated experiments, it may not be possible to do this so quickly. We need combinatoric rules (counting rules) to accomplish this.

- Combinatoric rules are used in probability to count the number of outcomes.

### 2.5 Tools for counting outcomes (sample points)

#### 2.5.1 Basic counting rule

**Basic counting rule (“$mn$ rule”):** Suppose we would like to count the number of paired outcomes formed by two factors. The first factor has $m$ outcomes. The second factor has $n$ outcomes. The total number of paired outcomes is $mn$.

**Example 2.11.** An experiment consists of rolling a die (with faces 1, 2, ..., 6) and tossing a coin (with sides H and T). The die has $m = 6$ outcomes. The coin has $n = 2$ outcomes. There are $mn = 12$ paired outcomes. The sample space for this experiment is

\[ S = \{(1, H), (2, H), (3, H), (4, H), (5, H), (6, H), (1, T), (2, T), (3, T), (4, T), (5, T), (6, T)\}. \]

There are $N = 12$ outcomes (sample points) in $S$. \(\square\)
Generalization: The basic counting rule can be generalized easily. Suppose there are \( k \) factors with

\[
\begin{align*}
    n_1 &= \text{number of outcomes for factor 1} \\
    n_2 &= \text{number of outcomes for factor 2} \\
    \vdots \\
    n_k &= \text{number of outcomes for factor } k.
\end{align*}
\]

The total number of outcomes is

\[\prod_{i=1}^{k} n_i = n_1 \times n_2 \times \cdots \times n_k.\]

**Example 2.12.** An experiment consists of selecting a standard South Carolina license plate which consists of 3 letters and 3 numbers. We can think of one outcome (sample point) in the underlying sample space \( S \) as having the following structure:

\[
(\quad \quad \quad \quad \quad \quad \quad \quad).
\]

**Q:** How many standard plates are possible; i.e., how many outcomes are in \( S \)?

**A:** There are

\[N = 26 \times 26 \times 26 \times 10 \times 10 \times 10 = 17576000\]

possible outcomes.

**Q:** Assume each outcome in \( S \) is equally likely (e.g., license plate letters/numbers are determined at random). What is the probability a randomly selected plate contains no repeat letters and no repeat numbers?

**A:** Define the event

\[A = \{\text{no repeat letters/numbers}\}.\]

The number of outcomes in \( A \) is

\[n_a = 26 \times 25 \times 24 \times 10 \times 9 \times 8 = 11232000.\]

Therefore,

\[P(A) = \frac{n_a}{N} = \frac{11232000}{17576000} \approx 0.6391. \square\]

**Example 2.13.** The birthday problem, revisited. An experiment consists of observing the birthday of \( M = 50 \) students. Assume 365 days. There are

\[N = 365 \times 365 \times 365 \times \cdots \times 365 = 365^{50}\]

possible outcomes. We can think of one outcome (sample point) in the underlying sample space \( S \) as having the following structure:

\[
(\quad \quad \quad \quad \quad \quad \quad \quad ).
\]
Q: Assume each outcome in $S$ is equally likely; e.g., no twins/triplets, etc. What is the probability there will be at least one shared birthday?

A: Define the event

$$A = \{\text{no shared birthdays}\}.$$ 

The number of outcomes in $A$ is

$$365 \times 364 \times 363 \cdots \times 317 \times 316 = 50! \left(\begin{array}{c}365 \\ 50 \end{array}\right).$$

Therefore,

$$P(A) = \frac{50! \left(\begin{array}{c}365 \\ 50 \end{array}\right)}{365^{50}} \approx 0.0296.$$ 

Using the complement rule, the probability of at least one shared birthday is

$$P(A) = 1 - P(\overline{A}) = 1 - \frac{50! \left(\begin{array}{c}365 \\ 50 \end{array}\right)}{365^{50}} \approx 0.9704.$$ 

Compare this with Example 2.3 where we used simulation to “estimate” this answer. □

2.5.2 Permutations

Remark: We have seen examples where constructing sample spaces requires us to work with distinct “objects;” e.g., license plate digits, students, etc. Counting the number of outcomes (in $S$ or in $A$) often requires us to count the number of ways distinct objects can be arranged in a sequence.

Terminology: A permutation is an arrangement of distinct objects in a particular order. Order is important.

Result: Suppose I have $n$ distinct objects. There are

$$n! = n(n-1)(n-2) \times \cdots \times 2 \times 1$$

ways to permute these objects (i.e., to arrange them in a particular order).

Example 2.14. Consider the experiment of arranging 10 distinct books on my bookshelf. There are

$$N = 10! = 3628800$$

possible permutations. We can think of one sample point in the underlying sample space $S$ as having the following structure:

$$\begin{array}{c} \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \end{array}.$$ 

Q: Assume each outcome in $S$ is equally likely. If there are 5 math books, 3 physics books, and 2 chemistry books, what is the probability that a randomly selected arrangement will keep like-subject books together?
A: Define the event

\[ A = \{ \text{like-subject books kept together} \} . \]

The number of outcomes in \( A \) can be found using the basic counting rule with

\[ n_1 = \text{number of ways to permute M, P, and C ordering} = 3! \]
\[ n_2 = \text{number of ways to permute M books} = 5! \]
\[ n_3 = \text{number of ways to permute P books} = 3! \]
\[ n_4 = \text{number of ways to permute C books} = 2! \]

The number of outcomes in \( A \) is

\[ n_a = 3! \times 5! \times 3! \times 2! = 8640. \]

Therefore,

\[ P(A) = \frac{n_a}{N} = \frac{8640}{3628800} \approx 0.0024. \]

Remark: In Example 2.14, our goal was to permute \( n \) distinct objects (i.e., books). In other problems, we first select \( r \) objects (from the available \( n \)) and then permute those only.

Result: From a collection of \( n \) distinct objects, we select and permute \( r \) of them \((r \leq n)\). The number of ways to do this is

\[ P^n_r = \frac{n!}{(n-r)!}. \]

The symbol \( P^n_r \) is read “the permutation of \( n \) things taken \( r \) at a time.”

Proof. Envision \( r \) slots. There are \( n \) ways to fill the first slot, \( n-1 \) ways to fill the second slot, and so on, until we get to the \( r \)th slot, where there are \( n-r+1 \) ways to fill it. From the basic counting rule, there are

\[ n(n-1)(n-2) \times \cdots \times (n-r+1) = \frac{n!}{(n-r)!} \]

different permutations.

Example 2.15. A personnel director for a corporation has hired 12 new engineers. She must pick 3 engineers to fill distinct positions (team leader, consultant, support staff member). Note that because these positions are inherently different, the selection ordering matters;

- e.g., the outcome (Jim, Mary, Celeste) and the outcome (Celeste, Jim, Mary) are different outcomes.

Conceptualize the selection of 3 engineers from 12 as a random experiment. We can think of one outcome (sample point) in the underlying sample space \( S \) as having the following structure:

\[
( \_ \_ \_ ).
\]
Because the ordering within outcomes is important, there are
\[ N = P^1_{12} = \frac{12!}{(12 - 3)!} = 12 \times 11 \times 10 = 1320 \]
outcomes in \( S \).

Q: Assume each outcome in \( S \) is equally likely. Suppose there are 6 engineers from USC and 6 from Clemson. What is the probability a USC graduate is selected as the team leader and the remaining 2 positions are filled by Clemson graduates?

A: Define the event
\[ A = \{ \text{USC team leader and Clemson graduates for other 2 positions} \} \]
The number of outcomes in \( A \) can be found using the basic counting rule with
\[
\begin{align*}
  n_1 &= \text{number of ways to select 1 USC graduate} = 6 \\
  n_2 &= \text{number of ways to select 2 Clemson graduates} = P^2_6
\end{align*}
\]
The number of outcomes in \( A \) is
\[ n_a = 6 \times P^6_2 = 6 \times 30 = 180. \]
Therefore,
\[ P(A) = \frac{n_a}{N} = \frac{180}{1320} \approx 0.1363. \]

2.5.3 Multinomial coefficients

Example 2.16. How many permutations of the letters in the word PEPPER are there?

Solution. Initially treat each of the 6 letters as distinct objects and emphasize this by writing
\[ P_1E_1P_2E_2P_3E_2R. \]
We know there are
\[ 6! = 720 \]
possible permutations of these 6 distinct objects. Now, because the letters in PEPPER really are not distinct, the number of possible permutations is smaller 720. By how much? Note that there are
\[
\begin{align*}
  3! \text{ ways to permute the Ps} \\
  2! \text{ ways to permute the Es} \\
  1! \text{ ways to permute the Rs.}
\end{align*}
\]
Therefore, \( 6! \) is \( 3! \times 2! \times 1! \) times too large. The number of permutations is
\[ \frac{6!}{3! \times 2! \times 1!} = 60. \]
Terminology: Multinomial coefficients arise in the algebraic expansion of the multinomial expression \((x_1 + x_2 + \cdots + x_k)^n\); i.e.,

\[
(x_1 + x_2 + \cdots + x_k)^n = \sum_{D} \binom{n}{n_1 n_2 \cdots n_k} x_1^{n_1} x_2^{n_2} \cdots x_k^{n_k},
\]

where \(D = \{(n_1, n_2, \ldots, n_k) : \sum_{i=1}^{k} n_i = n\}\). The multinomial coefficient

\[
\binom{n}{n_1 n_2 \cdots n_k} = \frac{n!}{n_1! n_2! \cdots n_k!}.
\]

Importance: In counting problems, multinomial coefficients are used to count

- the number of ways to permute \(n\) objects, of which \(n_1\) are “alike,” \(n_2\) are “alike,” and so on (see Example 2.16).

- the number of ways to partition \(n\) distinct objects into \(k\) distinct groups containing \(n_1, n_2, \ldots, n_k\) objects, respectively (where \(\sum_{i=1}^{k} n_i = n\)).

Example 2.17. A police department in a small city consists of 10 officers. The department policy is to have 5 officers patrolling the streets, 2 officers working at the station, and 3 officers on reserve. How many divisions (partitions) of the 10 officers are possible?

A: \(\binom{10}{5 \ 2 \ 3} = \frac{10!}{5! \ 2! \ 3!} = 2520\). □

Example 2.18. A signal is formed by arranging 9 flags in a line. There are 4 white flags, 3 blue flags, and 2 yellow flags. Envision the process of forming a signal as a random experiment.

We can think of one outcome (sample point) in the underlying sample space \(S\) as having the following structure:

\[
\text{( } \text{- } \text{- } \text{- } \text{- } \text{- } \text{- } \text{- } \text{- } \text{- } \text{ )}.
\]

Q: What is the probability the signal has the 4 white flags grouped together?

Note: We offer two solutions. The solutions differ in the way we conceptualize what a sample point looks like:

1. one solution treats flags of the same color as “indistinguishable” objects

2. one solution treats all 9 flags as distinct objects.

In the first conceptualization, a sample point might look like

\[
( B \ W \ W \ Y \ B \ Y \ B \ W \ W )
\]
In the second, a sample point might look like

\(( B_3 \ W_2 \ W_1 \ Y_1 \ B_1 \ Y_2 \ B_2 \ W_4 \ W_3 )\)

**Important:** Different counting rules are needed for each solution. In both solutions, we define

\( A = \{ \text{white flags grouped together} \} \)

and assume that outcomes (sample points) are equally likely.

**Solution 1:** Treat flags of the same color as “indistinguishable” objects. The number of sample points in \( S \) is \( N = \binom{9}{4 \ 3 \ 2} = \frac{9!}{4! \ 3! \ 2!} = 1260 \).

This is the number of ways to permute 9 objects, of which 4 are “alike,” 3 are “alike,” and 2 are “alike.” Now, we need to count the number of sample points in \( A \). We can do this using the basic counting rule:

\[
\begin{align*}
  n_1 &= \text{number of ways to select 4 adjacent positions for W flags} = 6 \\
  n_2 &= \text{number of ways to permute B/Y flags among the remaining positions} = \binom{5}{3 \ 2} \\
\end{align*}
\]

Therefore,

\[
  n_a = n_1 \times n_2 = 6 \times \binom{5}{3 \ 2} = 60
\]

and

\[
P(A) = \frac{n_a}{N} = \frac{60}{1260} \approx 0.0476.
\]

**Solution 2:** Treat all 9 flags as distinct objects. The number of sample points in \( S \) is \( N = 9! = 362880 \).

This is the number of ways to permute 9 distinct objects. Now, we need to count the number of sample points in \( A \). We can do this using the basic counting rule:

\[
\begin{align*}
  n_1 &= \text{number of ways to select 4 adjacent positions for W flags} = 6 \\
  n_2 &= \text{number of ways to permute W flags} = 4! \\
  n_3 &= \text{number of ways to permute B/Y flags among the remaining positions} = 5! \\
\end{align*}
\]

Therefore,

\[
  n_a = n_1 \times n_2 \times n_3 = 6 \times 4! \times 5! = 17280
\]

and

\[
P(A) = \frac{n_a}{N} = \frac{17280}{362880} \approx 0.0476. \quad \square
\]
2.5.4 Combinations

**Result:** From a collection of \( n \) distinct objects, we choose \( r \) of them \((r \leq n)\) without regard to the order in which the objects are chosen. The number of ways to do this is

\[
C_r^n = \binom{n}{r} = \frac{n!}{r! (n-r)!}.
\]

The symbol \( C_r^n \) is read “the combination of \( n \) things taken \( r \) at a time.”

**Remark:** To see why this makes sense, envision \( n \) distinct objects. The number of ways to partition these objects into 2 distinct groups, of which \( r \) are “alike” (i.e., the chosen objects) and \( n - r \) are “alike” (i.e., the objects not chosen) is given by the multinomial coefficient

\[
\binom{n}{r, n-r} = \frac{n!}{r! (n-r)!}.
\]

**Remark:** We adopt the notation \( \binom{n}{r} \), read “\( n \) choose \( r \),” henceforth as the symbol for \( C_r^n \). The terms \( \binom{n}{r} \) are called **binomial coefficients** because they arise in the algebraic expansion of a binomial; i.e.,

\[
(a + b)^n = \sum_{r=0}^{n} \binom{n}{r} a^{n-r} b^r.
\]

**Example 2.19.** In Example 2.15, a personnel director was tasked with choosing 3 engineers from 12 to fill distinct positions. If the positions are not distinct, then there are

\[
N = \binom{12}{3} = \frac{12!}{3! (12-3)!} = 220
\]

possible ways to select 3 engineers.

**Q:** Assume each combination is equally likely. Suppose there are 6 engineers from USC and 6 from Clemson. What is the probability of selecting 1 USC engineer and 2 from Clemson?

**A:** Define the event

\[
A = \{1 \text{ USC graduate and 2 Clemson graduates chosen}\}.
\]

The number of outcomes in \( A \) can be found using the basic counting rule with

\[
n_1 = \text{number of ways to select 1 USC graduate} = \binom{6}{1} = 6
\]

\[
n_2 = \text{number of ways to select 2 Clemson graduates} = \binom{6}{2} = 15
\]

The number of outcomes in \( A \) is

\[
n_a = n_1 \times n_2 = 6 \times 15 = 90.
\]

Therefore,

\[
P(A) = \frac{n_a}{N} = \frac{90}{220} \approx 0.4091.
\]
Remark: From Examples 2.15 and 2.19, one should note that, in general,

\[ P_{n,r} = r! \times \binom{n}{r}. \]

This formula highlights the difference between \( P_{n,r} \) and \( \binom{n}{r} \). To count the number of ways to permute \( n \) objects chosen \( r \) at a time, we first must choose the \( r \) objects. The binomial coefficient \( \binom{n}{r} \) counts the number of ways to do this. Then, once we have our \( r \) chosen objects, there are \( r! \) ways to permute them.

Example 2.20. Consider the experiment of drawing 5 cards from a standard deck of 52 cards (without replacement). We can conceptualize the sample space as

\[ S = \{[2S, 2D, 2H, 2C, 3S], [2S, 2D, 2H, 2C, 3D], [2S, 2D, 2H, 2C, 3H], \ldots, [A_S, A_D, A_H, A_C, K_C]\}. \]

The number of outcomes in \( S \) is

\[ N = \binom{52}{5} = \frac{52!}{5! (52-5)!} = 2598960. \]

Q: Assuming that each outcome in \( S \) is equally likely, what is the probability of getting “3 of a kind?”

A: Define the event

\[ A = \{ \text{“3 of a kind”} \}. \]

The number of outcomes in \( A \) can be found using the basic counting rule with

\[
\begin{align*}
    n_1 &= \text{number of ways to choose denomination} = \binom{13}{1} = 13 \\
    n_2 &= \text{number of ways to choose 3 suits} = \binom{4}{3} = 4 \\
    n_3 &= \text{number of ways to choose 2 other denominations} = \binom{12}{2} = 66 \\
    n_4 &= \text{number of ways to choose 1 card for each “other” denomination} = \left(\binom{4}{1}\right)^2 = 16
\end{align*}
\]

The number of outcomes in \( A \) is

\[ n_a = n_1 \times n_2 \times n_3 \times n_4 = 13 \times 4 \times 66 \times 16 = 54912. \]

Therefore,

\[ P(A) = \frac{n_a}{N} = \frac{54912}{2598960} \approx 0.0211. \]

Note: In choosing the 2 other denominations above (Step 3), it is important to remember that these 2 denominations must be different. If they are the same, then the hand is a “full house” instead (not a lesser “3 of a kind” hand).
Example 2.21. The matching problem, revisited. Suppose $M$ men are at a party, and each man is wearing a hat. Each man throws his hat into the center of the room. Each man then selects a hat at random. What is the probability at least one man selects his own hat; i.e., there is at least one “match”? Define

$$A = \{\text{at least one man selects his own hat}\}$$

and the events

$$A_i = \{\text{the } i\text{th man selects his own hat}\}, \quad i = 1, 2, \ldots, M,$$

so that

$$A = \bigcup_{i=1}^{M} A_i \implies P(A) = P\left(\bigcup_{i=1}^{M} A_i\right).$$

We now use the additive rule for $M$ events (see pp 11, notes). Note the following:

$$P(A_i) = \frac{(M-1)!}{M!} = \frac{1}{M} \quad \forall i = 1, 2, \ldots, M$$

$$P(A_{i_1} \cap A_{i_2}) = \frac{(M-2)!}{M!} \quad 1 \leq i_1 < i_2 \leq M$$

$$P(A_{i_1} \cap A_{i_2} \cap A_{i_3}) = \frac{(M-3)!}{M!} \quad 1 \leq i_1 < i_2 < i_3 \leq M$$

This pattern continues; the probability of the $M$-fold intersection is

$$P\left(\bigcap_{i=1}^{M} A_i\right) = \frac{(M-M)!}{M!} = \frac{1}{M!}.$$

Therefore, by the additive rule, we have

$$P\left(\bigcup_{i=1}^{M} A_i\right) = \sum_{i=1}^{M} P(A_i) - \sum_{i_1 < i_2} P(A_{i_1} \cap A_{i_2})$$

$$+ \sum_{i_1 < i_2 < i_3} P(A_{i_1} \cap A_{i_2} \cap A_{i_3}) - \cdots + (-1)^{M+1} P\left(\bigcap_{i=1}^{M} A_i\right)$$

$$= M \left(\frac{1}{M}\right) - \left(\frac{M}{2}\right) \frac{(M-2)!}{M!} + \left(\frac{M}{3}\right) \frac{(M-3)!}{M!} - \cdots + (-1)^{M+1} \frac{1}{M!}$$

$$= \sum_{k=1}^{M} (-1)^{k+1} \frac{M!}{k!} \frac{(M-k)!}{M!}$$

$$= \sum_{k=1}^{M} (-1)^{k+1} \frac{M!}{k!(M-k)!} \frac{(M-k)!}{M!} = 1 - \sum_{k=0}^{M} \frac{(-1)^k}{k!}.$$
2.6 Conditional probability and independence

**Remark:** The probability an event $A$ will often depend on other “related” events. If we know another one of these related events has occurred (or has not occurred), this may change the way we assess the likelihood of $A$ occurring.

**Example 2.22.** Consider the sample space in Example 2.10,

$$S = \{(M_1, M_2), (M_1, M_3), (M_1, W_1), (M_1, W_2), (M_2, M_3), (M_2, W_1), (M_2, W_2), (M_3, W_1), (M_3, W_2), (W_1, W_2)\},$$

where the experiment consisted of choosing two alternate jurors from three men and two women. Define the event

$$A = \{\text{two women are chosen}\}.$$

Assuming each outcome (sample point) in $S$ is equally likely, clearly

$$P(A) = \frac{n_a}{N} = \frac{1}{10}.$$

Now suppose we know that at least one of the jurors chosen is a woman. That is, the event

$$B = \{\text{at least one woman chosen}\} = \{(M_1, W_1), (M_1, W_2), (M_2, W_1), (M_2, W_2), (M_3, W_1), (M_3, W_2), (W_1, W_2)\}$$

has occurred. How does the knowledge of $B$ occurring influence how we assign probability to $A$?

In essence, a “new” sample space emerges when we know that $B$ has occurred, namely, the new sample space is $B$. Continuing to assume outcomes are equally likely, the probability $A$ occurs has now changed to

$$P(A|B) = \frac{1}{7}.$$

We write $P(A|B)$ to emphasize this is a conditional probability. □

**Terminology:** Let $A$ and $B$ be events in a sample space $S$. The **conditional probability** of $A$, given that $B$ has occurred, is given by

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

provided that $P(B) > 0$.

**Example 2.23.** Brazilian scientists have identified a new strain of the H1N1 virus. The genetic sequence of the new strain consists of alterations in the hemagglutinin protein, making it significantly different than the usual H1N1 strain. Public health officials wish to study the population of residents in Rio de Janeiro.
Suppose that in this population,

- the probability of catching the usual strain is 0.10
- the probability of catching the new strain is 0.05
- the probability of catching both strains is 0.01.

(a) Find the probability of catching the usual strain, given that the new strain is caught.
(b) Find the probability of catching the new strain, given that at least one strain is caught.

Solutions. Define the events

\[ A = \{ \text{resident catches usual strain} \} \]
\[ B = \{ \text{resident catches new strain} \}. \]

From the information above, we have \( P(A) = 0.10, P(B) = 0.05, \) and \( P(A \cap B) = 0.01. \)

(a) Using the definition of conditional probability,

\[ P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.01}{0.05} = 0.20. \]

(b) If “at least one strain is caught,” this means \( A \cup B \) has occurred. Therefore,

\[ P(B|A \cup B) = \frac{P(B \cap (A \cup B))}{P(A \cup B)} = \frac{P(B)}{P(A) + P(B) - P(A \cap B)} = \frac{0.05}{0.10 + 0.05 - 0.01} \approx 0.3571. \]

Note above that \( B \subset (A \cup B) \) so \( B \cap (A \cup B) = B. \)

Exercise: Find the probability of not catching the usual strain, given that the new strain is not caught. \( \Box \)

Important: Suppose \( P \) is a valid probability set function over \((S, \mathcal{B})\); i.e., it satisfies the Kolmogorov axioms. Provided that \( P(B) > 0, \) the conditional probability assignment

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

also satisfies the Kolmogorov axioms; i.e.,

1. \( P(A|B) \geq 0, \) for all \( A \in \mathcal{B} \)
2. \( P(B|B) = 1 \)
3. If \( A_1, A_2, ..., \in \mathcal{B} \) are pairwise mutually exclusive; i.e., \( A_i \cap A_j = \emptyset \; \forall i \neq j, \) then

\[ P \left( \bigcup_{i=1}^{\infty} A_i \bigg| B \right) = \sum_{i=1}^{\infty} P(A_i|B). \]
Implication: Because the way we assign conditional probability also satisfies the Kolmogorov axioms, all the probability rules we derived earlier have their respective “conditional versions.” For example,

1. Complement rule: $P(\overline{A}|B) = 1 - P(A|B)$
2. Monotonicity: If $A_1 \subset A_2$, then $P(A_1|B) \leq P(A_2|B)$
3. Additive rule: $P(A_1 \cup A_2|B) = P(A_1|B) + P(A_2|B) - P(A_1 \cap A_2|B)$.

Terminology: Suppose $A$ and $B$ are events in $S$. We say $A$ and $B$ are independent if

$$P(A \cap B) = P(A)P(B).$$

If both $P(A) > 0$ and $P(B) > 0$, then the following three conditions for independence are equivalent:

$$P(A \cap B) = P(A)P(B), \quad P(A|B) = P(A), \quad P(B|A) = P(B).$$

Remark: Do not confuse “independence” with what it means for $A$ and $B$ to be mutually exclusive. Independence means that the occurrence of $A$ does not affect whether $B$ occurs (and vice versa). If $A$ and $B$ are mutually exclusive, this means that $A$ and $B$ can not occur simultaneously.

Exercise: Suppose $P(A) > 0$ and $P(B) > 0$. Prove that if $A$ and $B$ are mutually independent, then $A$ and $B$ cannot be independent. Now go the other way. Prove that if $A$ and $B$ are independent, then $A$ and $B$ cannot be mutually exclusive.

Example 2.24. An electrical system consists of two components. The probability the second component functions in a satisfactory manner during its design life is 0.90. The probability at least one of the two components does so is 0.96. The probability both components do so is 0.75. Do the two components function independently?

Solution. Define the events

$$A = \{\text{component 1 functions}\}, \quad B = \{\text{component 2 functions}\}.$$ 

From the information above, we have $P(B) = 0.90$, $P(A \cup B) = 0.96$, and $P(A \cap B) = 0.75$. The additive rule gives

$$0.96 = P(A) + 0.90 - 0.75 \implies P(A) = 0.81.$$

However,

$$0.75 = P(A \cap B) \neq P(A)P(B) = 0.81(0.90) = 0.729.$$

Therefore, the events $A$ and $B$ are not independent.

Exercise: Check that $P(A|B) \neq P(A)$ and $P(B|A) \neq P(B)$. □
Result: Suppose $A$ and $B$ are events in $S$. If $A$ and $B$ are independent, then so are

(a) $A$ and $\overline{B}$
(b) $\overline{A}$ and $B$
(c) $\overline{A}$ and $\overline{B}$.

Proof. We prove part (a) only; the other parts follow similarly. Suppose $A$ and $B$ are independent. Then

$$P(\overline{A} \cap B) = P(\overline{A}|B)P(B) = [1 - P(A|B)]P(B) = [1 - P(A)]P(B) = P(\overline{A})P(B).$$

We used the fact that $A$ and $B$ were independent above when we wrote $P(A|B) = P(A)$. □

Multiplication rule: Suppose $A$ and $B$ are events in $S$. Then

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A).$$

This “rule” follows directly from the definition of conditional probability.

Generalization: Suppose $A_1, A_2, ..., A_n$ are events in $S$. Then

$$P\left(\bigcap_{i=1}^{n} A_i\right) = P(A_1) \times P(A_2|A_1) \times P(A_3|A_1 \cap A_2) \times \cdots \times P\left(A_n \bigg| \bigcap_{i=1}^{n-1} A_i\right).$$

Proof. We use mathematical induction. This is clearly true when $n = 2$ (see above). Assume the result holds for $n$ events. It suffices to show the induction step

$$P\left(\bigcap_{i=1}^{n+1} A_i\right) = P(A_1) \times P(A_2|A_1) \times P(A_3|A_1 \cap A_2) \times \cdots \times P\left(A_n \bigg| \bigcap_{i=1}^{n-1} A_i\right) \times P\left(A_{n+1} \bigg| \bigcap_{i=1}^{n} A_i\right).$$

Write

$$P\left(\bigcap_{i=1}^{n+1} A_i\right) = P\left(\bigcap_{i=1}^{n} A_i \cap A_{n+1}\right) = P\left(A_{n+1} \bigg| \bigcap_{i=1}^{n} A_i\right) P\left(\bigcap_{i=1}^{n} A_i\right).$$

However, note that

$$P\left(\bigcap_{i=1}^{n} A_i\right) = P(A_1) \times P(A_2|A_1) \times P(A_3|A_1 \cap A_2) \times \cdots \times P\left(A_n \bigg| \bigcap_{i=1}^{n-1} A_i\right)$$

is true by assumption. The result follows immediately. □

Discussion: The multiplication rule allows us to approach calculating the probability of an intersection “sequentially.” First, calculate $P(A_1)$ for the first event. Next, calculate $P(A_2|A_1)$ for the second event (given the first). Next, calculate $P(A_3|A_1 \cap A_2)$ for the third event (given the first two), and so on. The next example illustrates this approach.
Example 2.25. I am dealt a hand of 5 cards at random. What is the probability they are all spades?

Solution. Define the events

\[ A_i = \{ \text{the } i\text{th card is a spade} \}, \quad i = 1, 2, 3, 4, 5. \]

Assuming each card is randomly drawn from the deck,

\[
P(A_1) = \frac{13}{52},
\]

\[
P(A_2 | A_1) = \frac{12}{51},
\]

\[
P(A_3 | A_1 \cap A_2) = \frac{11}{50},
\]

\[
P(A_4 | A_1 \cap A_2 \cap A_3) = \frac{10}{49},
\]

\[
P(A_5 | A_1 \cap A_2 \cap A_3 \cap A_4) = \frac{9}{48}.
\]

Therefore, the probability all five cards are spades is

\[
P\left( \bigcap_{i=1}^{5} A_i \right) = P(A_1) \times P(A_2 | A_1) \times P(A_3 | A_1 \cap A_2) \times P(A_4 | A_1 \cap A_2 \cap A_3)
\]

\[
\times P(A_5 | A_1 \cap A_2 \cap A_3 \cap A_4)
\]

\[
= \frac{13}{52} \times \frac{12}{51} \times \frac{11}{50} \times \frac{10}{49} \times \frac{9}{48} \approx 0.0005.
\]

Remark: When I taught this class the last time, a student noted this calculation is easier if you simply regard the cards as belonging to two groups: spades and non-spades. There are \( \binom{13}{5} \) ways to draw 5 spades from 13. There are \( \binom{52}{5} \) possible hands. Thus, the probability of drawing 5 spades (assuming each hand is equally likely) is \( \frac{\binom{13}{5}}{\binom{52}{5}} \approx 0.0005. \)

Terminology: Suppose \( A_1, A_2, ..., A_n \) are events in \( S \). We say \( A_1, A_2, ..., A_n \) are mutually independent if for any sub-collection \( A_{i_1}, A_{i_2}, ..., A_{i_k} \), we have

\[
P\left( \bigcap_{j=1}^{k} A_{i_j} \right) = \prod_{j=1}^{k} P(A_{i_j}).
\]

Special case: Take 3 events \( A_1, A_2, \) and \( A_3 \). For these events to be mutually independent, we need them to be pairwise independent:

\[
P(A_1 \cap A_2) = P(A_1)P(A_2)
\]

\[
P(A_1 \cap A_3) = P(A_1)P(A_3)
\]

\[
P(A_2 \cap A_3) = P(A_2)P(A_3)
\]

and we also need

\[
P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3).
\]

For \( n > 2 \), mutual independence is a stronger condition than pairwise independence.
Exercise: Come up with an example of 3 events $A_1, A_2,$ and $A_3$ that are pairwise independent but not mutually independent. Hint: Think of rolling two fair dice with a sample space that regards all $N = 36$ outcomes as being equally likely.

Remark: Many random experiments can be envisioned as consisting of a sequence of $n$ “trials” that are viewed as independent (e.g., flipping a coin 10 times). If $A_i$ denotes the event associated with the $i$th trial, and the trials are mutually independent, then

$$P\left(\bigcap_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} P(A_i).$$

Example 2.26. The State Hygienic Laboratory at the University of Iowa tests thousands of residents for chlamydia every year. Suppose on a given day the lab tests $n = 30$ individual residents. Conceptualizing this as random experiment, the sample space can be written as

$$S = \{(0, 0, 0, ..., 0), (1, 0, 0, ..., 0), (0, 1, 0, ..., 0), ..., (1, 1, 1, ..., 1)\},$$

where “0” denotes a negative individual and “1” denotes a positive individual. Note that there are $N = 2^{30} = 1,073,741,824$ outcomes in $S$. However, these outcomes are probably not equally likely. Define the events

$$A_i = \{\text{$i$th individual is positive}\}, \ i = 1, 2, ..., 30.$$

Assume the 30 events $A_1, A_2, ..., A_{30}$ are mutually independent and that $P(A_i) = p$. What is the probability that at least one individual is positive?

Solution. First, note that by using the complement rule, we have $P(\overline{A}_i) = 1 - P(A_i) = 1 - p$, for $i = 1, 2, ..., 30$. Now, the event

$$A = \{\text{at least one individual is positive}\}$$

$$= \bigcup_{i=1}^{30} A_i.$$

The complement of $A$ is

$$\overline{A} = \{\text{all individuals are negative}\} = \bigcap_{i=1}^{30} \overline{A}_i$$

by DeMorgan’s Law. Because $A_1, A_2, ..., A_{30}$ are mutually independent (by assumption), the complements $\overline{A}_1, \overline{A}_2, ..., \overline{A}_{30}$ are also mutually independent. Therefore,

$$P(\overline{A}) = P\left(\bigcap_{i=1}^{30} \overline{A}_i\right) = \prod_{i=1}^{30} P(\overline{A}_i) = (1 - p)^{30}.$$ 

Finally,

$$P(A) = 1 - P(\overline{A}) = 1 - (1 - p)^{30}.$$

For example, if $p = 0.01$, then the probability of at least one positive individual among the 30 tested is $P(A) = 0.2603$. □
2.7 Law of Total Probability and Bayes’ Rule

**Law of Total Probability:** Suppose $A$ and $B$ are events in $S$. We can express $A$ as the union of two mutually exclusive events

$$A = (A \cap B) \cup (A \cap \overline{B}).$$

Therefore, by Axiom 3,

$$P(A) = P(A \cap B) + P(A \cap \overline{B})$$

$$= P(A|B)P(B) + P(A|\overline{B})P(\overline{B}).$$

This is called the **Law of Total Probability**.

**Remark:** The Law of Total Probability (LOTP) gives us a way to calculate $P(A)$ by relying instead on the conditional probabilities $P(A|B)$ and $P(A|\overline{B})$ and the (unconditional) probability of a related event $B$. More specifically, $P(A)$ is a linear combination of the conditional probabilities $P(A|B)$ and $P(A|\overline{B})$. The “weights” in the linear combination, $P(B)$ and $P(\overline{B})$, add to 1.

**Example 2.27.** An insurance company classifies drivers as “accident-prone” and “non-accident-prone.” The probability an accident-prone driver has an accident is 0.4. The probability a non-accident-prone driver has an accident is 0.2. The population is estimated to be 30 percent accident-prone.

(a) What is the probability that a policy-holder will have an accident?
(b) If a policy-holder has an accident, what is the probability that s/he was “accident-prone?”

**Solutions.** Define the events

$$A = \{\text{policy holder has an accident}\}$$

$$B = \{\text{policy holder is accident-prone}\}.$$

We are given $P(A|B) = 0.4$, $P(A|\overline{B}) = 0.2$, and $P(B) = 0.3$.

(a) We want to calculate $P(A)$. By the LOTP,

$$P(A) = P(A|B)P(B) + P(A|\overline{B})P(\overline{B})$$

$$= 0.4(0.3) + 0.2(0.7) = 0.26.$$

(b) We want to calculate $P(B|A)$. Note that

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

$$= \frac{0.4(0.3)}{0.26} \approx 0.462. \square$$
Note: From Example 2.27(b), we see that, in general,

\[ P(B|A) = \frac{P(A|B)P(B)}{P(A)} = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|\overline{B})P(\overline{B})}. \]

This is a special case of Bayes’ Rule.

Example 2.28. Diagnostic testing. A lab test is 95% effective at detecting a disease when it is present. It is 99% effective at declaring a subject negative when the subject is truly negative for the disease. Suppose 8% of the population has the disease.

(a) What is the probability a randomly selected subject will test positively?
(b) What is the probability a subject has the disease if his test is positive?

Solutions. Define the events

\[ D = \{\text{disease is present}\} \]
\[ A = \{\text{test is positive}\}. \]

We are given

\[ P(A|D) = 0.95 \quad (\text{“sensitivity”}) \]
\[ P(A|\overline{D}) = 0.99 \quad (\text{“specificity”}) \]
\[ P(D) = 0.08 \quad (\text{“prevalence”}). \]

(a) We want \( P(A) \). By the LOTP,

\[ P(A) = P(A|D)P(D) + P(A|\overline{D})P(\overline{D}) \]
\[ = 0.95(0.08) + 0.01(0.92) \approx 0.0852. \]

(b) We want \( P(D|A) \). By Bayes’ Rule,

\[ P(D|A) = \frac{P(A|D)P(D)}{P(A|D)P(D) + P(A|\overline{D})P(\overline{D})} \]
\[ = \frac{0.95(0.08)}{0.95(0.08) + 0.01(0.92)} \approx 0.892. \]

Remark: Bayes’ Rule allows us to “update” probabilities on the basis of observed information (in Example 2.28, this “observed information” is the test result):

<table>
<thead>
<tr>
<th>Prior probability</th>
<th>Test result</th>
<th>Posterior probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P(D) = 0.08 )</td>
<td>( A )</td>
<td>( P(D</td>
</tr>
<tr>
<td>( P(D) = 0.08 )</td>
<td>( \overline{A} )</td>
<td>( P(D</td>
</tr>
</tbody>
</table>

Note: \( P(D|A) \) in this example is called the “positive predictive value” (PPV). Calculate \( P(D|\overline{A}) \), the “negative predictive value” (NPV). □
Remark: For two events $A$ and $B$, the formulas for LOTP and Bayes’ Rule are given below:

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

Both of these formulas arise because the sample space $S$ can be written as $S = B \cup \overline{B}$, the union of two mutually exclusive events. In other words, the events $B$ and $\overline{B}$ “partition” the sample space. We now generalize LOTP and Bayes’ Rule for an arbitrary partition of $S$.

Terminology: A collection of events $B_1, B_2, ..., B_k$ forms a partition of the sample space $S$ if

$$\bigcup_{i=1}^{k} B_i = B_1 \cup B_2 \cup \cdots \cup B_k = S$$

and $B_i \cap B_j = \emptyset$, for $i \neq j$.

LOTP: Suppose $A$ is an event in $S$ and suppose $B_1, B_2, ..., B_k$ forms a partition of $S$. Then

$$P(A) = \sum_{i=1}^{k} P(A|B_i)P(B_i).$$

Proof. The event $A$ can be written as

$$A = A \cap S = A \cap \bigcup_{i=1}^{k} B_i = \bigcup_{i=1}^{k} (A \cap B_i).$$

Because $B_1, B_2, ..., B_k$ partition $S$, the events $A \cap B_1, A \cap B_2, ..., A \cap B_k$ are pairwise mutually exclusive. Therefore,

$$P(A) = P \left( \bigcup_{i=1}^{k} (A \cap B_i) \right) = \sum_{i=1}^{k} P(A \cap B_i) = \sum_{i=1}^{k} P(A|B_i)P(B_i),$$

the last step following from the multiplication rule. □

Bayes’ Rule: Suppose $A$ is an event in $S$ and suppose $B_1, B_2, ..., B_k$ forms a partition of $S$. Then

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^{k} P(A|B_i)P(B_i)}.$$

Proof. From the definition of conditional probability and the multiplication rule, note that

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

Now just write $P(A)$ out in its LOTP expansion. □
Example 2.29. For policy holders of a certain age, a life insurance company issues standard, preferred, and ultra-preferred policies. Among these policy holders,

- 60 percent are standard with a probability of 0.05 of dying next year.
- 30 percent are preferred with a probability of 0.03 of dying next year.
- 10 percent are ultra-preferred with a probability of 0.01 of dying next year.

(a) What is the probability a policy holder of this certain age dies next year?
(b) A policy holder of this certain age dies next year. What is the probability the deceased was a preferred policy holder?
(c) A policy holder of this certain age does not die next year. What is the probability this policy holder is an ultra-preferred policy holder?

Solutions. Define the events \( A = \{\text{policy holder dies next year}\} \) and

\[
B_1 = \{\text{policy holder has standard policy}\} \\
B_2 = \{\text{policy holder has preferred policy}\} \\
B_3 = \{\text{policy holder has ultra-preferred policy}\}.
\]

Note that \( \{B_1, B_2, B_3\} \) partition the sample space with \( P(B_1) = 0.60 \), \( P(B_2) = 0.30 \), and \( P(B_3) = 0.10 \). We are also given \( P(A|B_1) = 0.05 \), \( P(A|B_2) = 0.03 \), and \( P(A|B_3) = 0.01 \).

(a) We want \( P(A) \). By the LOTP,

\[
P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A|B_3)P(B_3) \\
= 0.05(0.60) + 0.03(0.30) + 0.01(0.10) = 0.04.
\]

(b) We want \( P(B_2|A) \). By Bayes’ Rule,

\[
P(B_2|A) = \frac{P(A|B_2)P(B_2)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A|B_3)P(B_3)} \\
= \frac{0.03(0.30)}{0.05(0.60) + 0.03(0.30) + 0.01(0.10)} = 0.225.
\]

Note how the “prior probability” \( P(B_2) = 0.30 \) has changed to \( P(B_2|A) = 0.225 \) when we learn that \( A \) has occurred.

(c) We want \( P(B_3|\overline{A}) \). By Bayes’ Rule,

\[
P(B_3|\overline{A}) = \frac{P(\overline{A}|B_3)P(B_3)}{P(\overline{A})} = \frac{(1 - P(A|B_3))P(B_3)}{1 - P(A)} = \frac{(1 - 0.01)(0.10)}{1 - 0.04} \approx 0.103.
\]

Note how the “prior probability” \( P(B_3) = 0.10 \) has changed to \( P(B_3|\overline{A}) \approx 0.103 \) when we learn that \( \overline{A} \) has occurred.
3 Discrete Random Variables and their Probability Distributions

3.1 Introduction

Recall: In Example 2.26 (pp 30, notes), we considered the problem of testing 30 Iowa residents for chlamydia. Conceptualizing this as random experiment, we wrote the sample space as

\[ S = \{(0, 0, 0, ..., 0), (1, 0, 0, ..., 0), (0, 1, 0, ..., 0), ..., (1, 1, 1, ..., 1)\}, \]

where “0” denotes a negative individual and “1” denotes a positive individual. Note that there are \( N = 2^{30} = 1,073,741,824 \) outcomes in \( S \).

Remark: Keeping track of outcomes in large unwieldy sample spaces like this is not practical. In this and other random experiments, it is much easier to reduce each outcome (sample point) to a numerical value.

Terminology: A random variable \( Y \) is a function whose domain is the sample space \( S \) and whose range is the set of real numbers \( \mathbb{R} = (-\infty, \infty) \). That is,

\[ Y : S \rightarrow \mathbb{R} \]

takes outcomes (sample points) in \( S \) and assigns them a real number.

Note: In the example above, define

\[ Y = \text{number of positives (out of 30)}. \]

Thinking of \( Y \) as a function, we see that, for example,

\[
\begin{align*}
Y((0, 0, 0, ..., 0)) &= 0 \\
Y((1, 0, 0, ..., 0)) &= 1 \\
Y((1, 1, 0, ..., 0)) &= 2 \\
Y((1, 1, 1, ..., 1)) &= 30.
\end{align*}
\]

The domain of \( Y \) is all 1,073,741,824 outcomes in \( S \). The range of \( Y \) is

\[ R = \{0, 1, 2, 3, ..., 30\}. \]

Terminology: The support of a random variable \( Y \) is the set of all possible values that \( Y \) can assume; i.e., it is the range of \( Y \) under the mapping \( Y : S \rightarrow \mathbb{R} \). We will denote the support by \( R \). It is understood that \( R \subseteq \mathbb{R} \).

Terminology: We call a random variable \( Y \) discrete if its support \( R \) is a finite or countable set. In other words, \( Y \) can assume a finite or (at most) a countable number of values.
Example 3.1. Consider the random experiment of rolling two dice and observing the face on each. The sample space for this experiment is 

\[ S = \{(1,1), (1,2), (1,3), (1,4), (1,5), (1,6), (2,1), (2,2), (2,3), (2,4), (2,5), (2,6), \\
(3,1), (3,2), (3,3), (3,4), (3,5), (3,6), (4,1), (4,2), (4,3), (4,4), (4,5), (4,6), \\
(5,1), (5,2), (5,3), (5,4), (5,5), (5,6), (6,1), (6,2), (6,3), (6,4), (6,5), (6,6)\}. \]

Assume the dice are “fair” so that each outcome (sample point) is equally likely; i.e., each outcome has probability \( \frac{1}{36} \).

Define the random variable 

\[ Y = \text{sum of the two faces}. \]

The support of \( Y \) is 

\[ R = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}. \]

Because \( R \) is a finite set, \( Y \) is a discrete random variable.

Q: How do we calculate probabilities like \( P(Y = 2) \)? like \( P(Y = 7) \)? like \( P(Y = 21) \)?

A: The “first principles” approach to doing this would be to find the inverse image of events like \( \{Y = 2\} \), \( \{Y = 7\} \), and \( \{Y = 21\} \) back on the original sample space \( S \) and then carry out the calculations there. For example,

\[ P(Y = 2) = P(\{(1,1)\}) = \frac{1}{36}. \]

Similarly,

\[ P(Y = 7) = P(\{(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)\}) = \frac{6}{36} \]

and

\[ P(Y = 21) = P(\emptyset) = 0. \]

Important: In general, the probability \( Y \) takes on the value \( y \), written \( P(Y = y) \), is the sum of the probabilities of the outcomes (sample points) in \( S \) that are assigned the value \( y \) under the mapping \( Y : S \rightarrow \mathbb{R} \). In notation,

\[ P(Y = y) = P(\{\omega \in S : Y(\omega) = y\}) = \sum_{\omega \in S \atop Y(\omega) = y} P(\{\omega\}), \]

where recall \( \omega \) denotes an outcome (sample point) in \( S \).

Terminology: The probability mass function (pmf) of a discrete random variable \( Y \) is the function defined by

\[ p_Y(y) = P(Y = y), \text{ for all } y. \]

If the value \( y \) is not in the support \( R \), then it is understood that \( p_Y(y) = 0 \). A discrete random variable’s pmf describes its probability distribution.
**Figure 3.1:** Probability mass function (pmf) of $Y$ in Example 3.1.

**Important:** The pmf of $Y$ in Example 3.1 (and in other examples) can be described by using a table, a graph, or a formula. In tabular form, we can write

<table>
<thead>
<tr>
<th>$y$</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
</table>

A graph of the pmf of $Y$ is shown in Figure 3.1 above. Finally, it is also possible to represent the pmf of $Y$ as a formula; i.e.,

$$p_Y(y) = \begin{cases} 
\frac{1}{36} (6 - |7 - y|), & y = 2, 3, ..., 12 \\
0, & \text{otherwise.}
\end{cases}$$

**Exercise:** In Example 3.1, find the pmf of

$$Y = \text{absolute difference of the two faces}.$$

For example, $Y((1, 1)) = |1 - 1| = 0$, $Y((1, 2)) = |1 - 2| = 1$, and so on. Depict your pmf in a table and a graph like above. □
Properties: The pmf of a discrete random variable $Y$ has the following properties:

1. $0 \leq p_Y(y) \leq 1$, for all $y$
2. the sum of the probabilities over all $y$ equals 1; i.e.,
   \[ \sum_{y \in R} p_Y(y) = 1. \]

These properties arise as a consequence of Axioms 1 and 2 (see pp 9, notes).

Example 3.2. I recently had a flight from Washington DC to Columbia. The plane had 66 seats on it and each seat was occupied; there were 36 females and 30 males on the flight. Suppose I selected 5 passengers at random and recorded $Y =$ number of males (out of 5).

Find the pmf of $Y$.

Solution. We can think of one outcome (sample point) in the underlying sample space $S$ as having the following structure:

\[
(\_\_\_\_\_). 
\]

For example, the outcomes

\[
(F_1 F_2 F_3 F_4 F_5) \quad \text{and} \quad (M_1 F_1 F_2 M_2 M_3),
\]

would produce the values $y = 0$ and $y = 3$, respectively. Note that there are

\[
N = \binom{66}{5} = 8936928
\]

outcomes in the sample space $S$ (the ordering of passenger selection doesn’t matter).

The pmf of $Y$ is the function $p_Y(y) = P(Y = y)$, which is nonzero when $y = 0, 1, 2, 3, 4, 5$. The number of outcomes in $S$ with $y$ males can be found by using the basic rule of counting:

\[
n_1 = \text{number of ways to select } y \text{ males from 30} = \binom{30}{y} \\

n_2 = \text{number of ways to select } 5 - y \text{ females from 36} = \binom{36}{5-y}
\]

Therefore, there are

\[
\binom{30}{y} \binom{36}{5-y}
\]

outcomes in $S$ with $y$ males. Assuming each outcome is equally likely,

\[
p_Y(y) = \begin{cases} 
\binom{30}{y} \binom{36}{5-y}, & y = 0, 1, 2, 3, 4, 5 \\
0, & \text{otherwise}
\end{cases}
\]
Here are the probabilities $p_Y(y) = P(Y = y)$ listed out in a table (to 3 dp):

<table>
<thead>
<tr>
<th>$y$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_Y(y)$</td>
<td>0.042</td>
<td>0.198</td>
<td>0.348</td>
<td>0.286</td>
<td>0.110</td>
<td>0.016</td>
</tr>
</tbody>
</table>

Note that these probabilities sum to 1, as required. A graph of the pmf of $Y$ is shown in Figure 3.2 above.

**Q:** What is the probability I select at least 4 males?

**A:** We can work directly from the pmf:

$$P(Y \geq 4) = P(Y = 4) + P(Y = 5) = p_Y(4) + p_Y(5) \approx 0.110 + 0.016 = 0.126. \square$$

This example illustrates the following general result.

**Result:** Suppose $Y$ is a **discrete** random variable with pmf $p_Y(y)$. The probability of an event $\{Y \in B\}$ is found by adding the probabilities $p_Y(y)$ for all $y \in B$; i.e.,

$$P(Y \in B) = \sum_{y \in B} p_Y(y).$$
Example 3.3. An experiment consists of rolling an unbiased die until the first “6” is observed. Let $Y$ denote the number of rolls needed. The pmf of $Y$ is given by

$$p_Y(y) = \begin{cases} \frac{1}{6} \left( \frac{5}{6} \right)^{y-1}, & y = 1, 2, 3, ... \\ 0, & \text{otherwise} \end{cases}$$

A graph of the pmf of $Y$ is shown in Figure 3.3 above.

Q: Is this a valid pmf?
A: Clearly, $0 \leq p_Y(y) \leq 1$, for each $y = 1, 2, 3, ...$. Do the probabilities $p_Y(y)$ sum to 1?

Recall: If $a \in \mathbb{R}$ and $|r| < 1$, then

$$\sum_{k=0}^{\infty} ar^k = \frac{a}{1 - r}.$$  

This is the formula for an infinite geometric sum. The condition $|r| < 1$ is needed or else the sum does not converge.
Note that
\[ \sum_{y=1}^{\infty} p_Y(y) = \sum_{y=1}^{\infty} \frac{1}{6} \left( \frac{5}{6} \right)^{y-1} = \sum_{k=0}^{\infty} \frac{1}{6} \left( \frac{5}{6} \right)^k = \frac{\frac{1}{6}}{1 - \frac{5}{6}} = 1. \]

Therefore, the pmf \( p_Y(y) \) is valid. \( \Box \)

**Example 3.4.** An insurance company models the number of claims per day, \( Y \), as a discrete random variable with pmf

\[
p_Y(y) = \begin{cases} 
\frac{1}{(y+1)(y+2)}, & y = 0, 1, 2, 3, ... \\
0, & \text{otherwise.}
\end{cases}
\]

A graph of the pmf of \( Y \) is shown in Figure 3.4 above.

**Q:** Is this a valid pmf?

**A:** Clearly, \( 0 \leq p_Y(y) \leq 1 \), for each \( y = 0, 1, 2, 3, ... \). Do the probabilities \( p_Y(y) \) sum to 1?

Note that we can rewrite
\[
\frac{1}{(y+1)(y+2)} = \frac{1}{y+1} - \frac{1}{y+2}.
\]
It follows that $\sum_{y=0}^{\infty} p_Y(y)$ is a telescoping sum; i.e.,

$$
\sum_{y=0}^{\infty} p_Y(y) = \sum_{y=0}^{\infty} \frac{1}{(y+1)(y+2)} = \sum_{y=0}^{\infty} \left( \frac{1}{y+1} - \frac{1}{y+2} \right)
$$

$$
= \left(1 - \frac{1}{2}\right) + \left(\frac{1}{2} - \frac{1}{3}\right) + \left(\frac{1}{3} - \frac{1}{4}\right) + \left(\frac{1}{4} - \frac{1}{5}\right) + \cdots = 1.
$$

Therefore, the pmf $p_Y(y)$ is valid. $\square$

### 3.2 Mathematical expectation

#### 3.2.1 Expected value

**Terminology:** Suppose $Y$ is a discrete random variable with pmf $p_Y(y)$ and support $R$. The **expected value** of $Y$ is

$$
E(Y) = \sum_{y \in R} yp_Y(y).
$$

In other words, $E(Y)$ is a weighted average of the possible values of $Y$. Each $y \in R$ is weighted by its corresponding probability $p_Y(y)$.

**Technical note:** If the support $R$ is countable but not finite, then $E(Y)$ may not exist. This occurs when the sum above does not converge absolutely. In other words, for $E(Y)$ to exist, we need

$$
\sum_{y \in R} |y| p_Y(y) < \infty.
$$

Of course, if $R$ is a finite set, then the sum $\sum_{y \in R} |y| p_Y(y)$ is finite and hence $E(Y)$ exists.

**Exercise:** Show that $E(Y)$ in Example 3.4 does not exist.

**Example 3.5.** Patient responses to a generic drug to control pain are scored on 5-point scale (1 = lowest pain level; 5 = highest pain level). In a certain population of patients, the pmf of the response $Y$ is given by

<table>
<thead>
<tr>
<th>$y$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_Y(y)$</td>
<td>0.38</td>
<td>0.27</td>
<td>0.18</td>
<td>0.11</td>
<td>0.06</td>
</tr>
</tbody>
</table>

A graph of the pmf of $Y$ is shown in Figure 3.5 (next page). The expected value of $Y$ is

$$
E(Y) = \sum_{y=1}^{5} yp_Y(y)
$$

$$
= 1(0.38) + 2(0.27) + 3(0.18) + 4(0.11) + 5(0.06) = 2.2.
$$
Interpretations: The expected value, or mean, of $Y$ can be interpreted in different ways:

- $E(Y)$ is the “center of gravity” on the pmf of $Y$ (see above). It’s located where the pmf would balance.

- $E(Y)$ is a “long run average.” In other words, if we observed the value of $Y$ over and over again (e.g., for a large number of patients in Example 3.5), then the average value would be close to $E(Y)$.

To illustrate this last interpretation, I used R’s sample function to sample 1000 values of $Y$ according to the pmf in Example 3.5:

```r
y = c(1,2,3,4,5)
prob = c(0.38,0.27,0.18,0.11,0.06)
sample.values = sample(y,1000,replace=TRUE,prob=prob)
> mean(sample.values)
[1] 2.203
```

The mean of these 1000 values was 2.203, which is very close to $E(Y) = 2.2$. □
Remark: In statistical applications, the expected value $E(Y)$ is called the population mean. This is the average value of $Y$ that would result from measuring every individual in the population (provided, of course, that we could and that the pmf of $Y$ was an accurate model for the population).

Example 3.6. An entomologist records $Y$, the number of insects that occupy a test plant. The pmf of $Y$ is given by

$$p_Y(y) = \begin{cases} 
\frac{e^{-1}}{y!}, & y = 0, 1, 2, 3, \ldots \\
0, & \text{otherwise}.
\end{cases}$$

A graph of the pmf of $Y$ is shown in Figure 3.6 above. Find $E(Y)$.

Solution. The expected value of $Y$ is

$$E(Y) = \sum_{y=0}^{\infty} y p_Y(y) = \sum_{y=0}^{\infty} y \frac{e^{-1}}{y!} = \sum_{y=1}^{\infty} \frac{e^{-1}}{y} = \sum_{y=1}^{\infty} \frac{e^{-1}}{(y-1)!} = e^{-1} \sum_{k=0}^{\infty} \frac{1}{k!} = e^{-1} e^1 = 1. \Box$$
Recall: The McLaurin series expansion of $f(x) = e^x$ is

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \cdots.$$ 

This expansion is valid for all $x \in \mathbb{R}$. For example, when $x = 1$, we have

$$e = e^1 = \sum_{k=0}^{\infty} \frac{1^k}{k!} = \sum_{k=0}^{\infty} \frac{1}{k!}.$$ 

In general, recall that the Taylor series expansion of the function $f(x)$ about the point $x = a$ is given by

$$f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(a)(x-a)^k}{k!}$$

$$= f(a) + f'(a)(x-a) + \frac{1}{2} f''(a)(x-a)^2 + \frac{1}{6} f^{(3)}(a)(x-a)^3 + \frac{1}{24} f^{(4)}(a)(x-a)^4 + \cdots.$$ 

A McLaurin series expansion is a Taylor series expansion when $a = 0$.

**Exercise:** Write out $f(x) = \ln(1+x)$, $f(x) = \cos x$, and $f(x) = 1/(1-x)$ in their McLaurin series expansions. Note for which values of $x \in \mathbb{R}$ the expansion is valid.

**Example 3.7.** *Discrete uniform distribution.* Suppose the random variable $Y$ has pmf

$$p_Y(y) = \begin{cases} \frac{1}{N}, & y = 1, 2, \ldots, N \\ 0, & \text{otherwise}, \end{cases}$$

where $N$ is a positive integer larger than 1. Find $E(Y)$.

**Solution.** The expected value of $Y$ is

$$E(Y) = \sum_{y=1}^{N} y \left( \frac{1}{N} \right) = \frac{1}{N} \sum_{y=1}^{N} y = \frac{1}{N} \left[ \frac{N(N+1)}{2} \right] = \frac{N+1}{2}.$$ 

Here, we have used the well known fact that the sum of the first $N$ positive integers; i.e.,

$$\sum_{y=1}^{N} y = 1 + 2 + 3 + \cdots + N = \frac{N(N+1)}{2}.$$ 

This can be proven using induction. If $N = 6$, then the discrete uniform distribution applies for the outcome of a fair die:

<table>
<thead>
<tr>
<th>$y$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_Y(y)$</td>
<td>1/6</td>
<td>1/6</td>
<td>1/6</td>
<td>1/6</td>
<td>1/6</td>
<td>1/6</td>
</tr>
</tbody>
</table>

The expected value of $Y$ is $E(Y) = (6+1)/2 = 3.5$. □
3.2.2 Functions of random variables

**Terminology**: Suppose $Y$ is a discrete random variable with pmf $p_Y(y)$ and support $R$. The **expected value** of $g(Y)$ is

$$E[g(Y)] = \sum_{y \in R} g(y)p_Y(y).$$

In other words, $E[g(Y)]$ is a weighted average of the possible values of $g(Y)$, where the probabilities $p_Y(y)$ play the role of the weights.

**Technical note**: If the support $R$ is countable but not finite, then $E[g(Y)]$ may not exist. This occurs when the sum above does not converge absolutely. In other words, for $E[g(Y)]$ to exist, we need

$$\sum_{y \in R} |g(y)|p_Y(y) < \infty.$$

**Example 3.8.** In Example 3.5, we used the pmf below to describe the population of patients’ responses to a generic drug to control pain:

<table>
<thead>
<tr>
<th>$y$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_Y(y)$</td>
<td>0.38</td>
<td>0.27</td>
<td>0.18</td>
<td>0.11</td>
<td>0.06</td>
</tr>
</tbody>
</table>

Find $E(Y^2)$, $E(\sqrt{Y})$, and $E(e^{tY})$, where $t$ is a constant.

**Solutions.** Using the result above, we have

$$E(Y^2) = 1^2(0.38) + 2^2(0.27) + 3^2(0.18) + 4^2(0.11) + 5^2(0.06) = 6.34$$

$$E(\sqrt{Y}) = \sqrt{1}(0.38) + \sqrt{2}(0.27) + \sqrt{3}(0.18) + \sqrt{4}(0.11) + \sqrt{5}(0.06) \approx 1.43$$

and

$$E(e^{tY}) = e^{t(1)}(0.38) + e^{t(2)}(0.27) + e^{t(3)}(0.18) + e^{t(4)}(0.11) + e^{t(5)}(0.06)$$

$$= 0.38e^t + 0.27e^{2t} + 0.18e^{3t} + 0.11e^{4t} + 0.06e^{5t}. \Box$$

**Properties**: In general, the expectation operator $E(\cdot)$ has certain properties. First, the expected value of a **constant** $c$ is $c$; i.e.,

$$E(c) = c.$$

This is easy to show when $Y$ is discrete with pmf $p_Y(y)$; note that

$$E(c) = \sum_{y \in R} cp_Y(y) = c \sum_{y \in R} p_Y(y) = c(1) = c.$$

Second, multiplicative constants can be moved outside the expectation; i.e.,

$$E[cg(Y)] = cE[g(Y)].$$
This is also easy to prove provided that \(E[g(Y)]\) exists; note that
\[
E[cg(Y)] = \sum_{y \in R} cg(y)p_Y(y) = c \sum_{y \in R} g(y)p_Y(y) = cE[g(Y)].
\]

Finally, taking expectations is additive; i.e.,
\[
E \left[ \sum_{j=1}^{k} g_j(Y) \right] = \sum_{j=1}^{k} E[g_j(Y)],
\]
provided that \(E[g_j(Y)]\) exists for each \(j = 1, 2, ..., k\). Note that
\[
E \left[ \sum_{j=1}^{k} g_j(Y) \right] = \sum_{y \in R} \sum_{j=1}^{k} g_j(y)p_Y(y)
= \sum_{y \in R} g_1(y)p_Y(y) + \sum_{y \in R} g_2(y)p_Y(y) + \cdots + \sum_{y \in R} g_k(y)p_Y(y)
= E[g_1(Y)] + E[g_2(Y)] + \cdots + E[g_k(Y)] = \sum_{j=1}^{k} E[g_j(Y)].
\]

These are called the **linearity properties** of the expectation.

**Another useful result:** If \(g(y) \geq 0\) for all \(y\), then \(E[g(Y)] \geq 0\). In other words, random variables that are nonnegative have nonnegative expectations.

### 3.2.3 Variance

**Terminology:** Suppose \(Y\) is a discrete random variable with mean \(E(Y) = \mu\). The **variance** of \(Y\) is
\[
\sigma^2 = V(Y) = E[(Y - \mu)^2] = \sum_{y \in R} (y - \mu)^2 p_Y(y).
\]

In other words, \(V(Y)\) is a weighted average of the possible values of \(g(Y) = (Y - \mu)^2\), where the probabilities \(p_Y(y)\) play the role of the weights.

**Note:** The variance \(V(Y)\) is the expected value of a “special” function of \(Y\), namely \(g(Y) = (Y - \mu)^2\). Similar technical requirements arise regarding existence; i.e., we need
\[
\sum_{y \in R} (y - \mu)^2 p_Y(y) < \infty
\]
for \(V(Y)\) to exist.

**Terminology:** The **standard deviation** of \(Y\) is the (positive) square root of the variance; i.e.,
\[
\sigma = \sqrt{\sigma^2} = \sqrt{V(Y)}.
\]
Example 3.9. In Example 3.5, we used the pmf below to describe the population of patients’ responses to a generic drug to control pain:

<table>
<thead>
<tr>
<th>y</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_y(y)$</td>
<td>0.38</td>
<td>0.27</td>
<td>0.18</td>
<td>0.11</td>
<td>0.06</td>
</tr>
</tbody>
</table>

Calculate $\sigma^2 = V(Y)$ and the standard deviation $\sigma$.

Solution. In Example 3.5, we calculated

$$E(Y) = \mu = 2.2.$$ 

Therefore, the variance of $Y$ is

$$\sigma^2 = \sum_{y=1}^{5} (y - \mu)^2 p_y(y)$$

$$= (1 - 2.2)^2(0.38) + (2 - 2.2)^2(0.27) + (3 - 2.2)^2(0.18)$$

$$+ (4 - 2.2)^2(0.11) + (5 - 2.2)^2(0.06) = 1.5.$$

The standard deviation of $Y$ is

$$\sigma = \sqrt{1.5} \approx 1.22.$$ 

Properties: The variance of a discrete random variable $Y$ has the following properties and interpretations:

1. The variance is nonnegative; i.e., $V(Y) \geq 0$. This is easy to see because

$$V(Y) = E[(Y - \mu)^2]$$

and $g(y) = (y - \mu)^2 \geq 0$ for all $y$; see the last result at the end of Section 3.2.2 (notes). Clearly $\sigma \geq 0$ as well.

2. Can $V(Y)$ ever be zero? It can, but only when all of the probability mass for $Y$ resides at one point, namely $y = \mu$. A random variable $Y$ with this property is called a degenerate random variable. Any constant $c$ can be thought of as a degenerate random variable.

3. Whereas the expected value $E(Y) = \mu$ measures the “center” or the “balance point” of a distribution, the variance $V(Y) = \sigma^2$ (and the standard deviation) measures the “spread” in the distribution. The larger $V(Y)$ is, the larger the spread.

4. The variance $V(Y) = \sigma^2$ is measured in the squared units of $Y$. The standard deviation $\sigma$ is measured in the same units as $Y$. Because of this, the standard deviation is easier for interpretation purposes.
**Variance computing formula:** Suppose $Y$ is a (discrete) random variable with mean $E(Y) = \mu$. The variance of $Y$ can be calculated as

$$V(Y) = E(Y^2) - [E(Y)]^2.$$ 

**Proof.** Using the definition of $V(Y)$, we have

$$V(Y) = E[(Y - \mu)^2] = E(Y^2 - 2\mu Y + \mu^2) = E(Y^2) - 2\mu E(Y) + \mu^2 = E(Y^2) - \mu^2 = E(Y^2) - [E(Y)]^2.$$ □

**Remark:** The variance computing formula is helpful because you only need to have $E(Y)$ and $E(Y^2)$ to find $V(Y)$. Note that, in general,

$$E(Y^2) \neq [E(Y)]^2.$$ 

The only time this is true is when $V(Y) = 0$; i.e., $Y$ is a degenerate random variable.

**Example 3.10.** *Discrete uniform distribution.* Suppose the random variable $Y$ has pmf

$$p_Y(y) = \begin{cases} \frac{1}{N}, & y = 1, 2, \ldots, N \\ 0, & \text{otherwise}, \end{cases}$$

where $N$ is a positive integer larger than 1. Find $V(Y)$.

**Solution.** We showed in Example 3.7 that

$$E(Y) = \frac{N + 1}{2}.$$ 

Therefore, we only need to find $E(Y^2)$. From the definition of expectation, we have

$$E(Y^2) = \sum_{y=1}^{N} y^2 \left( \frac{1}{N} \right) = \frac{1}{N} \sum_{y=1}^{N} y^2 = \frac{1}{N} \left[ \frac{N(N+1)(2N+1)}{6} \right] = \frac{(N+1)(2N+1)}{6}.$$ 

Here, we have used the well known fact that

$$\sum_{y=1}^{N} y^2 = 1^2 + 2^2 + 3^2 + \cdots + N^2 = \frac{N(N+1)(2N+1)}{6}.$$ 

Therefore, from the variance computing formula, we have

$$V(Y) = E(Y^2) - [E(Y)]^2 = \frac{(N+1)(2N+1)}{6} - \left( \frac{N+1}{2} \right)^2 = \frac{N^2 - 1}{12}. \square$$
Result: Suppose $Y$ is a (discrete) random variable and $a$ and $b$ are constants. Then

$$V(a + bY) = b^2V(Y).$$

Taking $b = 0$, we see that $V(a) = 0$ for any constant $a$. This makes sense intuitively. The variance is a measure of variability for a random variable; a constant (such as $a$) does not vary. Also, by taking $a = 0$, we see that $V(bY) = b^2V(Y)$.

### 3.3 Moment-generating functions

**Terminology:** The $k$th moment of a (discrete) random variable $Y$ is

$$\mu'_k = E(Y^k).$$

For example, the first four moments are

- $E(Y) = 1$st moment
- $E(Y^2) = 2$nd moment
- $E(Y^3) = 3$rd moment
- $E(Y^4) = 4$th moment

**Remark:** Note that the first moment $E(Y)$ is simply the expected value (or mean) or $Y$, which describes the “center” of the distribution of $Y$. Recall that

$$V(Y) = E(Y^2) - [E(Y)]^2$$

so the first two moments can be used to find $V(Y)$, which describes the “spread” in the distribution of $Y$.

**Terminology:** Suppose $Y$ is a discrete random variable with pmf $p_Y(y)$ and support $R$. The moment-generating function (mgf) of $Y$ is

$$m_Y(t) = E(e^{ty}) = \sum_{y \in R} e^{ty} p_Y(y),$$

provided this expectation is finite for all $t$ in an open neighborhood about $t = 0$; i.e., $\exists b > 0$ such that $E(e^{ty}) < \infty \forall t \in (-b, b)$. If no such $b > 0$ exists, then the moment generating function of $Y$ does not exist.

**Example 3.11.** In Example 3.6, we considered the discrete random variable $Y$ with pmf

$$p_Y(y) = \begin{cases} \frac{e^{-1}}{y!}, & y = 0, 1, 2, 3, \ldots \\ 0, & \text{otherwise}. \end{cases}$$

Find the mgf of $Y$. 

PAGE 50
Solution. The mgf of $Y$ is

$$m_Y(t) = E(e^{tY}) = \sum_{y=0}^{\infty} e^{ty} \frac{e^{-1}}{y!} = e^{-1} \sum_{y=0}^{\infty} \frac{(e^{t})^y}{y!} = e^{-1} \exp(e^{t}) = \exp(e^{t} - 1).$$

Above we used the fact that $\sum_{y=0}^{\infty} (e^{t})^y / y!$ is the McLaurin series expansion of $\exp(e^{t})$, which is a valid expansion for all $t \in \mathbb{R}$.

Example 3.12. Suppose $Y$ is a discrete random variable with pmf

$$p_Y(y) = \begin{cases} \left(\frac{1}{2}\right)^y, & y = 1, 2, 3, \ldots \\ 0, & \text{otherwise.} \end{cases}$$

Find the mgf of $Y$.

Solution. The mgf of $Y$ is

$$m_Y(t) = E(e^{tY}) = \sum_{y=1}^{\infty} e^{ty} \left(\frac{1}{2}\right)^y = \sum_{y=1}^{\infty} \left(\frac{e^{t}}{2}\right)^y = \left[ \sum_{y=0}^{\infty} \left(\frac{e^{t}}{2}\right)^y \right] - 1.$$

Note that the sum

$$\sum_{y=0}^{\infty} \left(\frac{e^{t}}{2}\right)^y$$

is an infinite geometric sum with common ratio

$$r = \frac{e^{t}}{2} < 1 \iff t < \ln 2.$$

Therefore, for all $t < \ln 2$, this sum converges and hence

$$m_Y(t) = \left[ \sum_{y=0}^{\infty} \left(\frac{e^{t}}{2}\right)^y \right] - 1 = \frac{1}{1 - \frac{e^{t}}{2}} - 1$$

$$= \frac{2}{2 - e^{t}} - 1 = \frac{e^{t}}{2 - e^{t}}.$$

Q: Why are mgfs useful?
A: Moment generating functions are functions that generate moments.

Important: If $Y$ is a random variable with mgf $m_Y(t)$, then

$$E(Y^k) = m_Y^{(k)}(0),$$

where

$$m_Y^{(k)}(0) = \frac{d^k}{dt^k} m_Y(t) \bigg|_{t=0}.$$ 

This shows how the moments of $Y$ can be found by differentiation. Note that derivatives are taken with respect to $t$. 

PAGE 51
Proof. Assume $Y$ is a discrete random variable with pmf $p_Y(y)$ and support $R$. For $k = 1$,
\[
\frac{d}{dt} m_Y(t) = \frac{d}{dt} \sum_{y \in R} e^{ty} p_Y(y) = \sum_{y \in R} \frac{d}{dt} e^{ty} p_Y(y) = \sum_{y \in R} ye^{ty} p_Y(y) = E(Y e^{tY}).
\]
Thus,
\[
\frac{d}{dt} m_Y(t) \bigg|_{t=0} = E(Y e^{tY}) \bigg|_{t=0} = E(Y).
\]
Taking higher-order derivatives, it follows that
\[
\frac{d^k}{dt^k} m_Y(t) \bigg|_{t=0} = E(Y^k),
\]
for any integer $k \geq 1$. □

Remark: In the argument above, we needed to assume that the interchange of the derivative and sum is justified. When the mgf exists, this interchange is justified.

Interesting: Writing $m_Y(t)$ in its McLaurin series expansion, we see that
\[
m_Y(t) = m_Y(0) + \frac{m_Y^{(1)}(0)}{1!}(t-0) + \frac{m_Y^{(2)}(0)}{2!}(t-0)^2 + \frac{m_Y^{(3)}(0)}{3!}(t-0)^3 + \cdots
\]
\[
= 1 + E(Y)t + \frac{E(Y^2)}{2}t^2 + \frac{E(Y^3)}{6}t^3 + \frac{E(Y^4)}{24}t^4 + \cdots
\]
\[
= \sum_{k=0}^{\infty} \frac{E(Y^k)}{k!} t^k.
\]
You can also see that
\[
E(Y^k) = \frac{d^k}{dt^k} m_Y(t) \bigg|_{t=0}
\]
by differentiating the RHS of $m_Y(t)$ written in its expansion (and evaluating at $t = 0$).

Example 3.13. Suppose $Y$ is a discrete random variable with pmf
\[
p_Y(y) = \begin{cases} 
(\frac{1}{2})^y, & y = 1, 2, 3, \ldots \\
0, & \text{otherwise.}
\end{cases}
\]
Find $E(Y)$ and $V(Y)$.

Solution. Using the definition of mathematical expectation, the first two moments of $Y$ are
\[
E(Y) = \sum_{y=1}^{\infty} y \left(\frac{1}{2}\right)^y,
\]
\[
E(Y^2) = \sum_{y=1}^{\infty} y^2 \left(\frac{1}{2}\right)^y.
\]
Neither of these sums are straightforward to calculate. Let’s use the mgf of $Y$ instead. Recall in Example 3.12 we found the mgf of $Y$ to be

$$m_Y(t) = \frac{e^t}{2 - e^t}.$$ 

The first two derivatives of $m_Y(t)$ are

$$\frac{d}{dt} m_Y(t) = \frac{2e^t}{(2 - e^t)^2},$$

$$\frac{d^2}{dt^2} m_Y(t) = \frac{2e^t(e^t + 2)}{(2 - e^t)^3}.$$ 

Therefore,

$$E(Y) = \frac{d}{dt} m_Y(t) \bigg|_{t=0} = \frac{2e^0}{(2 - e^0)^2} = 2.$$ 

The second moment is

$$E(Y^2) = \frac{d^2}{dt^2} m_Y(t) \bigg|_{t=0} = \frac{2e^0(e^0 + 2)}{(2 - e^0)^3} = 6.$$ 

Applying the variance computing formula, we have

$$V(Y) = E(Y^2) - [E(Y)]^2 = 6 - 4 = 2.$$ 

**Lesson:** In this example and elsewhere, finding $E(Y)$ and $E(Y^2)$ using the definition of mathematical expectation can be difficult. Using mgfs can be much easier. In other examples (e.g., Example 3.5, etc.), finding $E(Y)$ and $E(Y^2)$ using the definition of mathematical expectation is easy. There is no need to use mgfs in these examples.

### 3.4 Binomial distribution

**Important:** Many experiments consist of a sequence of “trials,” where

(i) each trial results in either a “success” or a “failure”

(ii) the probability of “success,” denoted by $p$, $0 < p < 1$, is the same on every trial

(iii) the trials are mutually independent.

Trials that obey these three properties are called **Bernoulli trials**.

**Terminology:** Let $Y$ denote the number of successes out of $n$ Bernoulli trials. Then $Y$ has a **binomial distribution** with parameters $n$ (the number of trials) and probability of success $p$. We write $Y \sim b(n, p)$. 

---

PAGE 53
Example 3.14. Consider each of the following situations involving a binomial random variable. Are you satisfied with the three Bernoulli trial assumptions in each case?

- I flip a coin \( n = 25 \) times and record \( Y \), the number of tails. If the coin is fair, then \( Y \sim b(n = 25, p = 0.5) \).

- In an agricultural study, it is determined that 40 percent of all plots respond to a certain treatment. Four plots are observed. If \( Y \) denotes the number of plots that respond to the treatment, then \( Y \sim b(n = 4, p = 0.4) \).

- In a biology experiment, 30 albino rats are injected with a drug that inhibits the synthesis of protein. The probability an individual rat will die from the drug before the study is complete is 0.15. If \( Y \) denotes the number of rats that die before the study is complete, then \( Y \sim b(n = 30, p = 0.15) \).

- Auditors estimate that 22 percent of insurance claims of a certain type are fraudulent. There are 189 claims this year. If \( Y \) denotes the number of fraudulent claims this year, then \( Y \sim b(n = 189, p = 0.22) \).

\[ \square \]

Note: Our goal is to derive the pmf of \( Y \sim b(n, p) \); i.e., to derive a formula for

\[ p_Y(y) = P(Y = y). \]

Among \( n \) Bernoulli trials, how can we get exactly \( y \) successes? We can think of one outcome (sample point) in the underlying sample space \( S \) as having the following structure

\[ ( \ldots S F F S S F S S F S \ldots ) \]

where each position (trial) is occupied by an S (for a “success”) or an F (for a “failure”). For example, if \( n = 10 \), one possible outcome looks like

\[ ( S F F S S F S S F S ) \]

and corresponds to \( y = 6 \) successes.

In general, any ordering of \( y \) successes (S’s) and \( n - y \) failures (F’s) occurs with probability

\[ p \times p \times \cdots \times p \times (1 - p) \times (1 - p) \times \cdots \times (1 - p) = p^y(1 - p)^{n-y}. \]

This is true because the trials are mutually independent and the probability of success (and the probability of failure) is the same on every trial. Thus, all we have to do is count the number of outcomes in the sample space with \( y \) successes; each one of these outcomes has the same probability \( p^y(1 - p)^{n-y} \). Counting this is the same as counting the number of ways to choose \( y \) positions (among the \( n \)) to contain a success S; there are \( \binom{n}{y} \) ways to do this.

PMF: The pmf of \( Y \sim b(n, p) \) is

\[ p_Y(y) = \begin{cases} \binom{n}{y} p^y(1 - p)^{n-y}, & y = 0, 1, 2, ..., n \\ 0, & \text{otherwise} \end{cases} \]
Recall: The binomial expansion of \((a + b)^n\) is given by

\[
(a + b)^n = \sum_{r=0}^{n} \binom{n}{r} a^{n-r} b^r.
\]

Q: Is the binomial pmf \(p_Y(y)\) valid?
A: Clearly, \(0 \leq p_Y(y) \leq 1\), for each \(y = 0, 1, 2, \ldots, n\). Do the probabilities \(p_Y(y)\) sum to 1? Letting \(a = 1 - p\), \(b = p\), and \(r = y\) in the binomial expansion formula above, we have

\[
[(1 - p) + p]^n = \sum_{y=0}^{n} \binom{n}{y} p^y (1 - p)^{n-y}.
\]

The LHS clearly equals 1. The RHS is the \(b(n, p)\) pmf. Thus, \(p_Y(y)\) is valid. □

MGF: The mgf of \(Y \sim b(n, p)\) is

\[
m_Y(t) = E(e^{tY}) = \sum_{y=0}^{n} e^{ty} \binom{n}{y} p^y (1 - p)^{n-y} = \sum_{y=0}^{n} \binom{n}{y} (pe^t)^y (1 - p)^{n-y} = (q + pe^t)^n,
\]

where \(q = 1 - p\). The last step follows from noting \(\sum_{y=0}^{n} \binom{n}{y} (pe^t)^y (1 - p)^{n-y}\) is the binomial expansion of \((q + pe^t)^n\). □

Mean/Variance: The mean and variance of \(Y \sim b(n, p)\) are

\[
E(Y) = np \\
V(Y) = np(1 - p).
\]

Proof. The first derivative of \(m_Y(t)\) with respect to \(t\) is

\[
m'_Y(t) = \frac{d}{dt} m_Y(t) = \frac{d}{dt} (q + pe^t)^n = n(q + pe^t)^{n-1} pe^t.
\]

Thus,

\[
E(Y) = \left. \frac{d}{dt} m_Y(t) \right|_{t=0} = n(q + pe^0)^{n-1} pe^0 = n(q + p)^{n-1} p = np,
\]

because \(q + p = 1\). To find \(V(Y)\), we can find the second moment \(E(Y^2)\) and then use the variance computing formula. The second derivative of \(m_Y(t)\) with respect to \(t\) is

\[
\frac{d^2}{dt^2} m_Y(t) = \frac{d}{dt} \binom{n}{y} (q + pe^t)^{n-1} pe^t = n(1)(q + pe^t)^{n-2}(pe^t)^2 + n(q + pe^t)^{n-1} pe^t.
\]

Thus,

\[
E(Y^2) = \left. \frac{d^2}{dt^2} m_Y(t) \right|_{t=0} = n(1)(q + pe^0)^{n-2}(pe^0)^2 + n(q + pe^0)^{n-1} pe^0 = n(n-1)p^2 + np.
\]

Finally,

\[
V(Y) = E(Y^2) - [E(Y)]^2 = n(n-1)p^2 + np - (np)^2 = np(1 - p). \square
\]
Note: WMS show how to derive $E(Y)$ by writing

$$E(Y) = \sum_{y=0}^{n} y \binom{n}{y} p^y (1 - p)^{n-y}$$

and then manipulating this sum (see pp 107-108). Calculating $E(Y^2)$ directly is difficult, so the authors instead find the second factorial moment

$$E[Y(Y - 1)] = \sum_{y=0}^{n} y(y-1) \binom{n}{y} p^y (1 - p)^{n-y}.$$

Note that

$$E[Y(Y - 1)] = E(Y^2 - Y) = E(Y^2) - E(Y) \implies E(Y^2) = E[Y(Y - 1)] + E(Y).$$

Factorial moments are discussed in Section 3.10 (pp 143-146, WMS).

**Example 3.15.** Physicians conjecture that 35 percent of renal cell carcinoma patients will respond positively to a new drug treatment. A small clinical trial tests the new drug in 30 patients. Let $Y$ denote the number of patients who will respond positively to the drug. If the Bernoulli trial assumptions hold for the patients (and the physicians’ conjecture is correct), then $Y \sim \text{b}(n = 30, p = 0.35)$. The pmf of $Y$ is shown in Figure 3.7 (next page).

Q: What is the probability exactly 10 patients respond positively? at most 10? at least 10?

A: We use the $\text{b}(n = 30, p = 0.35)$ pmf. The probability exactly 10 patients respond positively is

$$P(Y = 10) = p_Y(10) = \binom{30}{10} (0.35)^{10}(1 - 0.35)^{30-10} \approx 0.150.$$

The probability at most 10 patients respond positively is

$$P(Y \leq 10) = \sum_{y=0}^{10} \binom{30}{y} (0.35)^{y}(1 - 0.35)^{30-y} \approx 0.508.$$

The probability at least 10 patients respond positively is

$$P(Y \geq 10) = \sum_{y=10}^{30} \binom{30}{y} (0.35)^{y}(1 - 0.35)^{30-y} = 1 - \sum_{y=0}^{9} \binom{30}{y} (0.35)^{y}(1 - 0.35)^{30-y} \approx 0.642.$$

Here is the R code that will perform these calculations:

```R
> dbinom(10,30,0.35)
[1] 0.1502173
> pbinom(10,30,0.35)
[1] 0.5077582
> 1-pbinom(9,30,0.35)
[1] 0.6424591
```
Figure 3.7: Pmf of \( Y \sim b(n = 30, p = 0.35) \) in Example 3.15.

**Q:** What are \( E(Y) \) and \( V(Y) \)?

**A:** The mean of \( Y \) is

\[
E(Y) = np = 30(0.35) = 10.5 \text{ patients.}
\]

Therefore, we would expect 10.5 patients to respond positively. The variance of \( Y \) is

\[
V(Y) = np(1 - p) = 30(0.35)(1 - 0.35) = 6.825 \text{ (patients)}^2.
\]

The standard deviation is \( \sigma = \sqrt{6.825} \approx 2.61 \) patients. □

**Important:** In the \( b(n, p) \) family, when \( n = 1 \), the binomial pmf reduces to

\[
p_Y(y) = \left\{ \begin{array}{ll}
p^y(1-p)^{1-y}, & y = 0, 1 \\
0, & \text{otherwise.}
\end{array} \right.
\]

This is called the **Bernoulli distribution**. Shorthand notation is \( Y \sim b(1, p) \) or \( Y \sim \text{Bernoulli}(p) \). The Bernoulli distribution is used to model binary (0-1) outcomes; e.g., success/failure, agree/disagree, disease/healthy, etc.
3.5 Geometric distribution

Note: Recall the Bernoulli trial assumptions:

(i) each trial results in either a “success” or a “failure”
(ii) the probability of “success,” denoted by \( p \), \( 0 < p < 1 \), is the same on every trial
(iii) the trials are mutually independent.

Terminology: Suppose Bernoulli trials are continually observed. Let \( Y \) denote the number of trials to observe the first success. Then \( Y \) has a geometric distribution with probability of success \( p \). We write \( Y \sim \text{geom}(p) \).

PMF: The pmf of \( Y \sim \text{geom}(p) \) is

\[
p_Y(y) = \begin{cases} 
(1 - p)^{y-1}p, & y = 1, 2, 3, \ldots \\
0, & \text{otherwise.}
\end{cases}
\]

The form of this pmf makes sense; i.e., if the first success occurs on the \( y \)th trial, then the first \( y - 1 \) trials were failures. Each failure occurs with probability \( 1 - p \). The \( y \)th trial is a success (with probability \( p \)). Everything gets multiplied together because the Bernoulli trial outcomes are mutually independent.

Q: Is the geometric pmf \( p_Y(y) \) valid?

A: Clearly, \( 0 \leq p_Y(y) \leq 1 \), for each \( y = 1, 2, 3, \ldots \). Do the probabilities \( p_Y(y) \) sum to 1? We have

\[
\sum_{y=1}^{\infty} (1 - p)^{y-1}p = p \sum_{x=0}^{\infty} (1 - p)^x = \frac{p}{1 - (1 - p)} = 1.
\]

In the last step, we realize that \( \sum_{x=0}^{\infty} (1 - p)^x \) is an infinite geometric sum with common ratio \( 1 - p \).

MGF: The mgf of \( Y \sim \text{geom}(p) \) is

\[
M_Y(t) = E(e^{tY}) = \sum_{y=1}^{\infty} e^{ty}(1 - p)^{y-1}p = \frac{p}{q} \sum_{y=1}^{\infty} (qe^{t})^y
\]

\[
= \frac{p}{q} \left[ \sum_{y=0}^{\infty} (qe^{t})^y - 1 \right]
\]

\[
= \frac{p}{q} \left( \frac{1}{1 - qe^{t}} - 1 \right) = \frac{pe^{t}}{1 - qe^{t}},
\]

where \( q = 1 - p \). Note that the infinite geometric sum \( \sum_{y=0}^{\infty} (qe^{t})^y \) above converges and is equal to \( 1/(1 - qe^{t}) \) if and only if

\[
qe^{t} < 1 \iff t < -\ln q.
\]

Therefore, the mgf exists and is given by the formula above.
Mean/Variance: The mean and variance of \( Y \sim \text{geom}(p) \) are
\[
E(Y) = \frac{1}{p} \\
V(Y) = \frac{q}{p^2},
\]
where \( q = 1 - p \).

Proof. The first derivative of \( m_Y(t) \) with respect to \( t \) is
\[
m_Y'(t) = \frac{d}{dt} m_Y(t) = \frac{d}{dt} \left( \frac{pe^t}{1 - qe^t} \right) = \frac{pe^t(1 - qe^t) - pe^t(-qe^t)}{(1 - qe^t)^2} = \frac{pe^t}{(1 - qe^t)^2}.
\]
Therefore,
\[
E(Y) = \left. \frac{d}{dt} m_Y(t) \right|_{t=0} = \frac{pe^0}{(1 - qe^0)^2} = \frac{p}{(1 - q)^2} = \frac{1}{p}.
\]
To find \( V(Y) \), we can find the second moment \( E(Y^2) \) and then use the variance computing formula. The second derivative of \( m_Y(t) \) with respect to \( t \) is
\[
\frac{d^2}{dt^2} m_Y(t) = \frac{pe^t(1 - qe^t)^2 - 2pe^t(1 - qe^t)(-qe^t)}{(1 - qe^t)^4}.
\]
Therefore,
\[
E(Y^2) = \left. \frac{d^2}{dt^2} m_Y(t) \right|_{t=0} = \frac{pe^0(1 - qe^0)^2 - 2pe^0(1 - qe^0)(-qe^0)}{(1 - qe^0)^4} = \frac{p^3 + 2p^2q}{p^4} = \frac{p + 2q}{p^2}.
\]
Finally,
\[
V(Y) = E(Y^2) - [E(Y)]^2 = \frac{p + 2q}{p^2} - \left( \frac{1}{p} \right)^2 = \frac{q}{p^2}.
\]

Note: WMS show how to derive \( E(Y) \) and \( V(Y) \) directly by writing
\[
E(Y) = \sum_{y=1}^{\infty} y(1 - p)^{y-1}p \\
E[Y(Y - 1)] = \sum_{y=1}^{\infty} y(y - 1)(1 - p)^{y-1}p
\]
and then manipulating these sums; see pp 116-117.

Example 3.16. An EPA engineer is tasked with observing water specimens from lakes in northeast Georgia. In this region, each specimen has a 20 percent chance of containing a particular organic pollutant. Let \( Y \) denote the number of specimens observed to find the first one containing the pollutant. If the Bernoulli trial assumptions hold for the specimens, then \( Y \sim \text{geom}(p = 0.20) \). The pmf of \( Y \) is shown in Figure 3.8 (next page).
Here are the first few probabilities:

\[
\begin{align*}
P(Y = 1) &= p_Y(1) = (1 - 0.20)^{1-1}(0.20) = 0.20 \\
P(Y = 2) &= p_Y(2) = (1 - 0.20)^{2-1}(0.20) = 0.16 \\
P(Y = 3) &= p_Y(3) = (1 - 0.20)^{3-1}(0.20) = 0.128 \\
P(Y = 4) &= p_Y(4) = (1 - 0.20)^{4-1}(0.20) = 0.1024 \\
P(Y = 5) &= p_Y(5) = (1 - 0.20)^{5-1}(0.20) = 0.08192.
\end{align*}
\]

The probability the first water specimen containing the pollutant is observed among the first five specimens is

\[
P(Y \leq 5) = P(Y = 1) + P(Y = 2) + P(Y = 3) + P(Y = 4) + P(Y = 5)
= \sum_{y=1}^{5} (1 - 0.20)^{y-1}(0.20) = 0.67232. \quad \square
\]

> `pgeom(5-1,0.20)`
> [1] 0.67232
3.6 Negative binomial distribution

Note: Recall the Bernoulli trial assumptions:

(i) each trial results in either a “success” or a “failure”
(ii) the probability of “success,” denoted by $p$, $0 < p < 1$, is the same on every trial
(iii) the trials are mutually independent.

Terminology: Suppose Bernoulli trials are continually observed. Let $Y$ denote the number of trials to observe the $r$th success, where $r \geq 1$. Then $Y$ has a negative binomial distribution with waiting parameter $r$ and probability of success $p$. We write $Y \sim \text{nib}(r, p)$.

Note: When $r = 1$, the $\text{nib}(r, p)$ distribution reduces to the $\text{geom}(p)$ distribution. We can think of the negative binomial distribution as a generalization of the geometric; i.e., where one is “waiting” for more successes.

PMF: The pmf of $Y \sim \text{nib}(r, p)$ is

$$p_Y(y) = \begin{cases} \frac{(y-1)^{r-1}(1-p)^{y-r}}{(r-1)(y-r)}, & y = r, r+1, r+2, \ldots \\
0, & \text{otherwise}. \end{cases}$$

The form of this pmf can be explained intuitively. If the $r$th success occurs on the $y$th trial, then $r-1$ successes must have occurred during the first $y-1$ trials. The number of sample points (in the underlying sample space) where this occurs is $\binom{y-1}{r-1}$, which counts the number of ways one can choose the locations of $r-1$ successes among the 1st $y-1$ trials. Because the trials are mutually independent, the probability of each of these sample points is $p^{r-1}(1-p)^{y-r}$. Therefore, the probability of exactly $r-1$ successes among the first $y-1$ trials is $\binom{y-1}{r-1} p^{r-1}(1-p)^{y-r}$. On the $y$th trial, we observe the $r$th success (this occurs with probability $p$). Because the $y$th trial is independent of the previous $y-1$ trials, we have

$$P(Y = y) = \underbrace{\binom{y-1}{r-1} p^{r-1}(1-p)^{y-r}}_{\text{pertains to 1st } y-1 \text{ trials}} \times p = \binom{y-1}{r-1} p^{r}(1-p)^{y-r}.$$

MGF: The mgf of $Y \sim \text{nib}(r, p)$ is

$$\left( \frac{pe^t}{1-qe^t} \right)^r,$$

for $t < -\ln q$, where $q = 1-p$.

Note: When $r = 1$, the $\text{nib}(r, p)$ mgf reduces to the $\text{geom}(p)$ mgf. Interesting!
Proof. The mgf of $Y \sim \text{nib}(r, p)$ is

$$M_Y(t) = E(e^{ty}) = \sum_{y=r}^{\infty} e^{ty} \left( \frac{y-1}{r-1} \right) p^r (1-p)^{y-r} = (pe^t)^r \sum_{y=r}^{\infty} \left( \frac{y-1}{r-1} \right) (qe^t)^{y-r} = \left( \frac{pe^t}{1-qe^t} \right)^r,$$

for $1 - qe^t > 0 \iff t < -\ln q$. That $\sum_{y=r}^{\infty} \left( \frac{y-1}{r-1} \right) (qe^t)^{y-r} = (1 - qe^t)^{-r}$ follows from the lemma below. □

**Lemma.** Suppose $r$ is a nonnegative integer. Then

$$\sum_{y=r}^{\infty} \left( \frac{y-1}{r-1} \right) (qe^t)^{y-r} = (1 - qe^t)^{-r}.$$

**Proof.** Consider the function $f(w) = (1 - w)^{-r}$, where $r$ is a nonnegative integer. It is easy to show that

$$f'(w) = r(1 - w)^{-(r+1)},$$
$$f''(w) = r(r+1)(1 - w)^{-(r+2)},$$
$$f'''(w) = r(r+1)(r+2)(1 - w)^{-(r+3)},$$

and so on. In general, $f^{(z)}(w) = r(r+1) \cdots (r+z-1)(1 - w)^{-(r+z)}$, where $f^{(z)}(w)$ denotes the $z$th derivative of $f$ with respect to $w$. Note that

$$f^{(z)}(w) \bigg|_{w=0} = r(r+1) \cdots (r+z-1).$$

Now writing $f(w)$ in its McLaurin Series expansion, we have

$$f(w) = \sum_{z=0}^{\infty} \frac{f^{(z)}(0)}{z!} w^z = \sum_{z=0}^{\infty} \frac{r(r+1) \cdots (r+z-1)}{z!} w^z = \sum_{z=0}^{\infty} \left( \frac{r+z-1}{r-1} \right) w^z.$$

Letting $w = qe^t$ and $z = y - r$ proves the lemma. □

**Mean/Variance:** The mean and variance of $Y \sim \text{nib}(r, p)$ are

$$E(Y) = \frac{r}{p}, $$
$$V(Y) = \frac{rq}{p^2},$$

where $q = 1 - p$. Note again these formulae for $E(Y)$ and $V(Y)$ reduce to those for the geometric distribution when $r = 1$.

**Proof.** Exercise. □
Example 3.17. At an automotive plant, 15 percent of all paint batches sent to the lab for chemical analysis do not conform to specifications. Let $Y$ denote the number of batches to find the 4th one that does not conform. If the Bernoulli trial assumptions hold for the batches, then $Y \sim \text{nib}(r = 4, p = 0.15)$. The pmf of $Y$ is shown in Figure 3.9 (above).

Q: What is the probability no more than three nonconforming batches will be observed among the first 30 batches sent to the lab?

A: This will occur when the fourth nonconforming batch is observed on the 31st batch sent to the lab, the 32nd, the 33rd, etc. Therefore,

$$P(Y \geq 31) = 1 - P(Y \leq 30)$$

$$= 1 - \sum_{y=4}^{30} \binom{y-1}{4-1} (0.15)^4 (0.85)^{y-4} \approx 0.322. \square$$

> 1-pnbinom(30-4,4,0.15)

[1] 0.3216599
3.7 Hypergeometric distribution

**Setting:** Consider a population of $N$ objects and suppose each object belongs to one of two dichotomous classes: Class 1 or Class 2. For example, the objects and classes might be

- Poker chips: red/blue
- People: diseased/healthy
- Plots of land: respond to treatment/not.

In the population of interest, we have

- $N = \text{total number of objects}$
- $r = \text{number of objects in Class 1}$
- $N - r = \text{number of objects in Class 2}$.

We sample $n$ objects from the population at random and without replacement. Define

$Y = \text{number of objects in Class 1 (among the } n \text{ sampled)}.$

Then $Y$ has a **hypergeometric distribution** with population size $N$, sample size $n$, and number of Class 1 objects $r$. We write $Y \sim \text{hyper}(N, n, r)$.

**Remark:** We have already seen an example of this distribution in Example 3.2 (notes). In this example, the “objects” were passengers and the classes were male/female; i.e.,

- $N = \text{total number of passengers} = 66$
- $r = \text{number of males} = 30$
- $N - r = \text{number of females} = 36$.

We sampled $n = 5$ passengers at random and without replacement from the population of 66 passengers and recorded $Y$, the number of males among those sampled. In this example, $Y \sim \text{hyper}(N = 66, n = 5, r = 30)$. By conceptualizing the selection of $n = 5$ passengers as a random experiment, we derived the pmf of $Y$ to be

$$p_Y(y) = \begin{cases} \frac{\binom{r}{y} \binom{N-r}{n-y}}{\binom{N}{n}}, & y = 0, 1, 2, 3, 4, 5 \\ 0, & \text{otherwise.} \end{cases}$$

The hypergeometric pmf derivation generalizes immediately.

**PMF:** The pmf of $Y \sim \text{hyper}(N, n, r)$ is

$$p_Y(y) = \begin{cases} \frac{\binom{r}{y} \binom{N-r}{n-y}}{\binom{N}{n}}, & y = 0, 1, 2, ..., n \\ 0, & \text{otherwise.} \end{cases}$$
Comparison: The motivation for the hypergeometric distribution should remind us of the underlying framework for the binomial; i.e., we record the number of Class 1 objects ("successes") out of \( n \) ("trials"). The difference here is that

- the population size \( N \) is finite
- sampling is done without replacement.

To understand further, suppose

\[
p = \frac{r}{N} = \text{proportion of Class 1 objects in the population.}
\]

Because sampling from the population is done without replacement, the value of \( p \) changes from trial to trial. This violates the Bernoulli trial assumptions, so technically the binomial model does not apply. However, one can show mathematically that

\[
\lim_{N \to \infty} \frac{r}{N} \to p
\]

This result implies that if the population size \( N \) is "large," the \( \text{hyper}(N, n, r) \) distribution and the \( b(n, p = r/N) \) distribution should be very close to each other even when one samples without replacement. Of course, if one samples from a population with replacement, then \( p = r/N \) remains fixed and hence the binomial model applies regardless of how large \( N \) is.

Example 3.18. A supplier ships parts to a company in lots of 1000 parts. Suppose a lot contains 100 defective parts and 900 non-defective parts. An operator selects 10 parts at random and without replacement. What is the probability he selects no more than 2 defective parts?

Hypergeometric: Because sampling is done without replacement, a hypergeometric model applies. We recognize

\[
N = \text{total number of parts} = 1000
\]
\[
r = \text{number of defectives} = 100
\]
\[
N - r = \text{number of non-defectives} = 900.
\]

Let \( Y \) denote the number of defective parts (i.e., "Class 1 objects") out of \( n = 10 \). Then \( Y \sim \text{hyper}(N = 1000, n = 10, r = 100) \) and

\[
P(Y \leq 2) = P(Y = 0) + P(Y = 1) + P(Y = 2)
\]
\[
= \left( \frac{100}{0} \right) \left( \frac{900}{10} \right) + \left( \frac{100}{1} \right) \left( \frac{900}{9} \right) + \left( \frac{100}{2} \right) \left( \frac{900}{8} \right)
\]
\[
\approx 0.3469 + 0.3894 + 0.1945 = 0.9308.
\]

\[> \text{phyper}(2,100,900,10)\]

[1] 0.9307629
Figure 3.10: Example 3.18. Left: Pmf of $Y \sim \text{hyper}(N = 1000, n = 10, r = 100)$. Right: Pmf of $Y \sim b(n = 10, p = 0.10)$.

**Binomial**: The population proportion of defective parts is

$$p = \frac{100}{1000} = 0.10.$$  

Therefore, the $b(n = 10, p = 0.10)$ model should offer a good approximation to the (exact) answer obtained from the hypergeometric calculation. We have

$$P(Y \leq 2) = P(Y = 0) + P(Y = 1) + P(Y = 2)$$

$$= \binom{10}{0}(0.10)^0(0.90)^{10} + \binom{10}{1}(0.10)^1(0.90)^9 + \binom{10}{2}(0.10)^2(0.90)^8$$

$$\approx 0.3487 + 0.3874 + 0.1937 = 0.9298.$$  

> pbinom(2,10,0.10)

[1] 0.9298092

Figure 3.10 (above) shows the hypergeometric and binomial pmfs used in this problem. Note that they are nearly identical in appearance. □

**Q:** Is the hypergeometric pmf $p_Y(y)$ valid?

**A:** Clearly, $0 \leq p_Y(y) \leq 1$, for each $y = 1, 2, 3, \ldots$. Do the probabilities $p_Y(y)$ sum to 1? The answer is yes, of course, but showing this is not trivial. It suffices to show

$$\sum_{y=0}^{n} \binom{r}{y} \binom{N-r}{n-y} = \binom{N}{n}.$$  

See Exercise 3.216 (pp 156, WMS).
Mean/Variance: The mean and variance of \( Y \sim \text{hyper}(N, n, r) \) are

\[
E(Y) = n \left( \frac{r}{N} \right) \\
V(Y) = n \left( \frac{r}{N} \right) \left( \frac{N-r}{N} \right) \left( \frac{N-n}{N-1} \right) 
\]

Deriving these formulas is not trivial either. The mgf of \( Y \sim \text{hyper}(N, n, r) \) exists, but its form is not very friendly. Therefore, to derive \( E(Y) \), we will have to appeal directly to the definition of expected value; note that

\[
E(Y) = \sum_{y=0}^{n} yp_Y(y) = \sum_{y=0}^{n} y \frac{r}{n} \left( \frac{N-r}{n} \right) \left( \frac{N-n}{n-1} \right) = \sum_{y=1}^{n} y \frac{r}{n} \left( \frac{N-r}{n} \right) \left( \frac{N-n}{n} \right)
\]

The denominator in the pmf of \( Y \) can be written as

\[
\begin{align*}
\left( \begin{array}{c} N \\ n \end{array} \right) &= \frac{N!}{n!(N-n)!} = \frac{N}{n} \left( \frac{(N-1)!}{(n-1)!(N-n)!} \right) = \frac{N}{n} \left( \frac{N-1}{n-1} \right)
\end{align*}
\]

Therefore,

\[
E(Y) = \frac{n}{N} \sum_{y=1}^{n} y \frac{r}{n} \left( \frac{N-r}{n} \right) \left( \frac{N-n}{n-1} \right) = \frac{n}{N} \sum_{y=1}^{n} y \frac{r}{n} \frac{(r-1)!}{(r-y)!(n-y)!(N-r-n+y)!} \left( \frac{N-r}{n-1} \right)
\]

\[
= \frac{nnr}{N} \sum_{y=1}^{n} \frac{(r-1)!}{(r-y)!(n-y)!(N-r-n+y)!} \left( \frac{N-r}{n-1} \right)
\]

\[
| \text{if } x = y | - 1 \frac{nnr}{N} \sum_{x=0}^{n-1} \frac{(r-1)}{x!(r-1-x)!} \left( \frac{N-r}{n-1-x} \right) \left( \frac{N-n}{n-1} \right)
\]

However,

\[
\sum_{x=0}^{n-1} \frac{(r-1)}{x!(r-1-x)!} \frac{N-r}{n-1-x} \frac{N-n}{n-1} = 1
\]

because the summand is the pmf of \( X \sim \text{hyper}(N-1, n-1, r-1) \) and we sum over the support of this random variable; i.e., \( x = 0, 1, ..., n-1 \). Thus, the result. \( \Box \)

Note: To derive \( V(Y) \), it is easier to first calculate the second factorial moment

\[
E[Y(Y-1)]=\sum_{y=0}^{n} y(y-1) \frac{r}{n} \left( \frac{N-r}{n} \right) \left( \frac{N-n}{n} \right)
\]

Recall that

\[
E[Y(Y-1)] = E(Y^2 - Y) = E(Y^2) - E(Y) \implies E(Y^2) = E[Y(Y-1)] + E(Y).
\]
Interesting: If the population size $N \to \infty$ so that $r/N \to p \in (0, 1)$, note that

$$E(Y) = n \left( \frac{r}{N} \right) \to np,$$

the mean of the $b(n, p)$ distribution. Similarly,

$$V(Y) = n \left( \frac{r}{N} \right) \left( \frac{N-r}{N} \right) \left( \frac{N-n}{N-1} \right) \to np(1-p),$$

which is the variance of the $b(n, p)$ distribution. Neither result is surprising given the result on pp 65 (notes); i.e., if the $hyper(N, n, r)$ pmf converges to the $b(n, p)$ pmf as $N \to \infty$ and $r/N \to p$, then the corresponding moments should converge as well.

### 3.8 Poisson distribution

**Setting:** Suppose we count the number of “occurrences” in a continuous interval of time (or space). A Poisson process enjoys the following properties:

1. the number of occurrences in non-overlapping intervals are independent random variables
2. the probability of an occurrence in a sufficiently short interval is proportional to the length of the interval
3. the probability of 2 or more occurrences in a sufficiently short interval is zero.

Suppose a counting process satisfies the three conditions above. Define

$$Y = \text{the number of occurrences in a unit interval of time (or space)}.$$

Our goal is to find an expression for $p_Y(y) = P(Y = y)$, the pmf of $Y$.

**Derivation:** Partition the unit interval $[0, 1]$ into $n$ subintervals, each of size $1/n$.

- If $n$ is sufficiently large (i.e., much larger than $y$), then we can approximate the probability $y$ events occur in the unit interval by finding the probability that exactly one event (occurrence) occurs in exactly $y$ of the subintervals.
- By Property (2), we know that the probability of one event in any one subinterval is proportional to the subinterval’s length, say $\lambda/n$, where $\lambda$ is the proportionality constant.
- By Property (3), the probability of more than one occurrence in any subinterval is zero (for $n$ large).
Consider the occurrence/non-occurrence of an event in each subinterval as a Bernoulli trial. By Property (1), we have a sequence of \( n \) Bernoulli trials, each with probability of “success” \( p = \frac{\lambda}{n} \). Thus, a binomial (approximate) calculation gives

\[
P(Y = y) \approx \binom{n}{y} \left( \frac{\lambda}{n} \right)^y \left( 1 - \frac{\lambda}{n} \right)^{n-y}.
\]

To improve the approximation for \( P(Y = y) \), we let \( n \) grow large without bound; i.e., let \( n \to \infty \). We have

\[
\lim_{n \to \infty} P(Y = y) = \lim_{n \to \infty} n^n \frac{n!}{y!(n-y)!} \left( \frac{1}{n} \right)^y \left( 1 - \frac{\lambda}{n} \right)^{n-y} = \lim_{n \to \infty} \frac{n(n-1) \cdots (n-y+1)}{n^y} \frac{\lambda^y}{y!} \left( 1 - \frac{\lambda}{n} \right)^y = \lim_{n \to \infty} \frac{\lambda^y e^{-\lambda}}{y!}.
\]

Now, the limit of the product is the product of the limits:

\[
\lim_{n \to \infty} a_n = \lim_{n \to \infty} \frac{n(n-1) \cdots (n-y+1)}{n^y} = 1, \quad \lim_{n \to \infty} b_n = \lim_{n \to \infty} \frac{\lambda^y}{y!} = \frac{\lambda^y}{y!}, \quad \lim_{n \to \infty} c_n = \lim_{n \to \infty} \left( 1 - \frac{\lambda}{n} \right)^n = e^{-\lambda}, \quad \lim_{n \to \infty} d_n = \lim_{n \to \infty} \left( \frac{1}{1 - \frac{\lambda}{n}} \right)^y = 1.
\]

We have shown that

\[
\lim_{n \to \infty} P(Y = y) = \frac{\lambda^y e^{-\lambda}}{y!}.
\]

We say that \( Y \) follows a Poisson distribution with parameter \( \lambda \). Shorthand notation is \( Y \sim \text{Poisson}(\lambda) \).

**PMF:** The pmf of \( Y \sim \text{Poisson}(\lambda) \) is

\[
p_Y(y) = \begin{cases} \frac{\lambda^y e^{-\lambda}}{y!}, & y = 0, 1, 2, \\ 0, & \text{otherwise.} \end{cases}
\]

**Q:** Is the Poisson pmf \( p_Y(y) \) valid?

**A:** Clearly, \( 0 \leq p_Y(y) \leq 1 \), for each \( y = 0, 1, 2, \ldots \). Do the probabilities \( p_Y(y) \) sum to 1? We have

\[
\sum_{y=0}^{\infty} \frac{\lambda^y e^{-\lambda}}{y!} = e^{-\lambda} \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} = e^{-\lambda} e^\lambda = 1.
\]

Recall that \( \sum_{y=0}^{\infty} \lambda^y / y! \) is the McLaurin series expansion of \( e^\lambda \). \( \square \)
The mgf of $Y \sim \text{Poisson}(\lambda)$ is
\[
m_Y(t) = E(e^{tY}) = \sum_{y=0}^{\infty} e^{ty} \frac{\lambda^y e^{-\lambda}}{y!} = e^{-\lambda} \sum_{y=0}^{\infty} \frac{(\lambda e^t)^y}{y!} = e^{-\lambda} e^{\lambda e^t} = \exp[\lambda(e^t - 1)]. \]

**Mean/Variance:** The mean and variance of $Y \sim \text{Poisson}(\lambda)$ are
\[
E(Y) = \lambda \\
V(Y) = \lambda.
\]

Proof. The first derivative of $m_Y(t)$ with respect to $t$ is
\[
m'_Y(t) = \frac{d}{dt}m_Y(t) = \frac{d}{dt} \exp[\lambda(e^t - 1)] = \lambda e^t \exp[\lambda(e^t - 1)].
\]

Thus,
\[
E(Y) = \left. \frac{d}{dt}m_Y(t) \right|_{t=0} = \lambda e^0 \exp[\lambda(e^0 - 1)] = \lambda.
\]

To find $V(Y)$, we can find the second moment $E(Y^2)$ and then use the variance computing formula. The second derivative of $m_Y(t)$ with respect to $t$ is
\[
\frac{d^2}{dt^2} m_Y(t) = \frac{d}{dt} \left( \lambda e^t \exp[\lambda(e^t - 1)] \right) = \lambda e^t \exp[\lambda(e^t - 1)] + (\lambda e^t)^2 \exp[\lambda(e^t - 1)].
\]

Thus,
\[
E(Y^2) = \left. \frac{d^2}{dt^2} m_Y(t) \right|_{t=0} = \lambda e^0 \exp[\lambda(e^0 - 1)] + (\lambda e^0)^2 \exp[\lambda(e^0 - 1)] = \lambda + \lambda^2.
\]

Finally,
\[
V(Y) = E(Y^2) - [E(Y)]^2 = \lambda + \lambda^2 - \lambda^2 = \lambda. \]

**Note:** WMS show how to derive $E(Y)$ directly by writing
\[
E(Y) = \sum_{y=0}^{\infty} \frac{\lambda^y e^{-\lambda}}{y!}
\]
and then manipulating this sum. This is easy to do. Note that
\[
E(Y) = \sum_{y=0}^{\infty} \frac{\lambda^y e^{-\lambda}}{y!} = \sum_{y=1}^{\infty} \frac{\lambda^y e^{-\lambda}}{y!} = \lambda \sum_{y=1}^{\infty} \frac{\lambda^{y-1} e^{-\lambda}}{(y-1)!}.
\]
Letting $x = y - 1$ in the last sum, we get
\[
E(Y) = \lambda \sum_{x=0}^{\infty} \frac{\lambda^x e^{-\lambda}}{x!} = \lambda.
\]
Figure 3.11: Pmf of $Y \sim \text{Poisson}(\lambda = 1.5)$ in Example 3.19.

To derive $V(Y)$, we could calculate the second factorial moment $E[Y(Y - 1)]$ and then use the fact that

$$E[Y(Y - 1)] = E(Y^2) - E(Y) \implies E(Y^2) = E[Y(Y - 1)] + E(Y).$$

However, in this case, it is just as easy to calculate $E(Y^2)$ directly. Note that

$$E(Y^2) = \sum_{y=0}^{\infty} y^2 \frac{\lambda^y e^{-\lambda}}{y!} = \sum_{y=1}^{\infty} y^2 \frac{\lambda^y e^{-\lambda}}{y!} = \lambda \sum_{y=1}^{\infty} \frac{\lambda^{y-1} e^{-\lambda}}{(y-1)!}.$$

Letting $x = y - 1$ in the last sum, we get

$$E(Y^2) = \lambda \sum_{x=0}^{\infty} (x + 1) \frac{\lambda^x e^{-\lambda}}{x!} = \lambda E(X + 1),$$

where $X \sim \text{Poisson}(\lambda)$. Therefore, $E(Y^2) = \lambda(\lambda + 1) = \lambda^2 + \lambda$, which is the same as what we got by finding $E(Y^2)$ using the mgf of $Y$.

**Example 3.19.** In a certain region in the northeast US, the number of severe weather events per year $Y$ is assumed to have a Poisson distribution with mean $\lambda = 1.5$. The pmf of $Y \sim \text{Poisson}(\lambda = 1.5)$ is shown in Figure 3.11 above.
Q: What is the probability there are four or more severe weather events in a given year?
A: We want to find \( P(Y \geq 4) \). Work directly with the Poisson pmf; first note that
\[
P(Y \leq 3) = P(Y = 0) + P(Y = 1) + P(Y = 2) + P(Y = 3) = \frac{(1.5)^0 e^{-1.5}}{0!} + \frac{(1.5)^1 e^{-1.5}}{1!} + \frac{(1.5)^2 e^{-1.5}}{2!} + \frac{(1.5)^3 e^{-1.5}}{3!} \\
\approx 0.223 + 0.335 + 0.251 + 0.126 = 0.935.
\]
By the complement rule,
\[
P(Y \geq 4) = 1 - P(Y \leq 3) \approx 1 - 0.935 = 0.065.
\]

\[
> 1-\text{ppois}(3,1.5) \[1\] 0.06564245
\]

Q: A company buys a policy to insure its revenue in the event of severe weather that shuts down business. The policy pays nothing for the first such weather event of the year and $10,000 for each one thereafter, until the end of the year. Calculate the expected amount paid to the company under this policy during a one-year period.
A: First note that if \( Y = 0 \) or \( Y = 1 \), then the company receives nothing according to the policy. It is only when there are 2 or more severe weather events does a payout occur, and this payout is $10,000 for each event. Therefore, the payout when viewed as a function of \( Y \) is given by
\[
g(Y) = \begin{cases} 0, & Y = 0, 1 \\ 10000(Y - 1), & Y = 2, 3, 4, ...
\end{cases}
\]
and we want to calculate \( E[g(Y)] \). From the definition of mathematical expectation, we have
\[
E[g(Y)] = \sum_{y=0}^{\infty} g(y) \frac{(1.5)^y e^{-1.5}}{y!}
\]
\[
= 0 \times \frac{(1.5)^0 e^{-1.5}}{0!} + 0 \times \frac{(1.5)^1 e^{-1.5}}{1!} + \sum_{y=2}^{\infty} 10000(y - 1) \frac{(1.5)^y e^{-1.5}}{y!}
\]
\[
= 10000 \left[ \sum_{y=0}^{\infty} (y - 1) \frac{(1.5)^y e^{-1.5}}{y!} - (1 - 1) \times \frac{(1.5)^1 e^{-1.5}}{1!} - (0 - 1) \times \frac{(1.5)^0 e^{-1.5}}{0!} \right].
\]
Note that
\[
\sum_{y=0}^{\infty} (y - 1) \frac{(1.5)^y e^{-1.5}}{y!} = E(Y - 1) = E(Y) - 1 = 1.5 - 1 = 0.5.
\]
Therefore,
\[
E[g(Y)] = 10000(0.5 - 0 + e^{-1.5}) \approx 7231.30.
\]
The expected payout to the company during a one-year period is $7,231.30. \( \square \)