Introduction to Probability Theory

Nathaniel E. Helwig

University of Minnesota

1 Experiments and Events

The field of "probability theory" is a branch of mathematics that is concerned with describing the likelihood of different outcomes from uncertain processes. Probability theory is the cornerstone of the field of Statistics, which is concerned with assessing the uncertainty of inferences drawn from random samples of data. Thus, we need to understand basics of probability theory to comprehend some of the basic principles used in inferential statistics. Before defining what the word "probability" means, I will introduce some terminology to motivate the need for thinking probabilistically.

Definition. A simple experiment is some action that leads to the occurrence of a single outcome s from a set of possible outcomes S. Note that the single outcome s is referred to as a sample point, and the set of possible outcomes S is referred to as the sample space.

***Example* 1.** Suppose that you flip a coin $n \geq 2$ times and record the number of times you observe a "heads". The sample space is $S = \{0, 1, \ldots, n\}$, where $s = 0$ corresponds to observing no heads and $s = n$ corresponds to observing only heads.

***Example* 2.** Suppose that you pick a card at random from a standard deck of 52 playing cards. The sample points are the individual cards in the deck (e.g., the Queen of Spades is one possible sample point), and the sample space is the collection of all 52 cards.

***Example* 3.** Suppose that you roll two standard (six-sided) dice and sum the obtained numbers. The sample space is $S = \{2, 3, \ldots, 11, 12\}$, where $s = 2$ corresponds to rolling "snake eyes" (i.e., two 1's) and $s = 12$ corresponds to rolling "boxcars" (i.e., two 6's).

Definition. An event A refers to any possible subspace of the sample space S, i.e., $A \subseteq S$, and an elementary event is an event that contains a single sample point s.

***Example* 4.** For the coin flipping example, we could define the events

- $A = \{0\}$ (i.e., we observe no heads)
- $B = \{1, 2\}$ (i.e., we observe 1 or 2 heads)
- $C = \{c \mid c \text{ is an even number}\}$ (i.e., we observe an even number of heads)

***Example* 5.** For the playing card example, we could define the events

- $A = \{\text{Queen of Spades}\}$ (i.e., we draw the Queen of Spades)
- $B = \{b \mid b \text{ is a Queen}\}$ (i.e., we draw a card that is a Queen)
- $C = \{c \mid c \text{ is a Spade}\}$ (i.e., we draw a card that is a Spade)

***Example* 6.** For the dice rolling example, we could define the events

- $A = \{2\}$ (i.e., we roll snake eyes)
- $B = \{7, 11\}$ (i.e., we roll natural or yo-leven)
- $C = \{c \mid c \text{ is an even number}\}$ (i.e., we roll dice that sum to an even number)

For each of the above examples, A is an elementary event, whereas B and C are not elementary events. Note that this is assuming that 0 is considered an even number, which ensures that C is a non-elementary event when there are only $n = 2$ coin flips.

Definition. A sure event is an event that always occurs, and an impossible event (or null event) is an event that never occurs.

***Example* 7.** For the coin flipping example, $E = \{e \mid e \text{ is an integer satisfying } 0 \leq e \leq n\}$ is a sure event and $I = \{i \mid i > n\}$ is an impossible event.

***Example* 8.** For the playing card example, $E = \{e \mid e \text{ is a Club, Diamond, Heart, or Spade}\}$ is a sure event and $I = \{\text{Joker}\}$ is an impossible event.

***Example* 9.** For the dice rolling example, $E = \{e \mid e \text{ is an integer satisfying } 2 \leq e \leq 12\}$ is a sure event and $I = \{i \mid i > 12\}$ is an impossible event.

Definition. Two events A and B are said to be mutually exclusive if $A \cap B = \emptyset$, i.e., if one event occurs, then the other event can not occur. Two events A and B are said to be exhaustive if $A \cup B = S$, i.e., if one of the two events must occur.

***Example* 10.** For the coin flipping example, the two events $A = \{0\}$ and $B = \{n\}$ are mutually exclusive events, whereas $A = \{a \mid a \text{ is an even number}\}$ and $B = \{b \mid b \text{ is an odd number}\}$ are exhaustive events.

***Example* 11.** For the playing card example, the two events $A = \{a \mid a \text{ is a Spade}\}$ and $B = \{b \mid b \text{ is a Club}\}$ are mutually exclusive events, whereas $A = \{a \mid a \text{ is a Club or Spade}\}$ and $B = \{b \mid b \text{ is a Diamond or Heart}\}$ are exhaustive events.

***Example* 12.** For the dice rolling example, the two events $A = \{2\}$ and $B = \{12\}$ are mutually exclusive events, whereas $A = \{a \mid a \text{ is an even number}\}$ and $B = \{b \mid b \text{ is an odd number}\}$ are exhaustive events.

2 What is a Probability?

Definition. A probability is a real number (between 0 and 1) that we assign to events in a sample space to represent their likelihood of occurrence. The notation $P(A)$ denotes the probability of the event $A \subseteq S$.

There are two differing perspectives on how to interpret what a probability actually means:[1]

- The "physical" interpretation views a probability as the relative frequency of events that would occur in the *long run*, i.e., if the simple experiment were repeated a very large number of times. This interpretation is used in Frequentist statistical inference.
- The "evidential" interpretation views a probability as means of representing the subjective plausibility of a statement, regardless of whether any random process is involved. This interpretation is used in Bayesian statistical inference.

We will use the "physical" interpretation, given that this course is focused on Frequentist statistical inference. But there is some merit to the "evidential" interpretation of probability in a variety of real-world applications (because the *long run* isn't always relevant).

[1] For a discussion, see `https://en.wikipedia.org/wiki/Probability_interpretations`

3 Axioms of Probability

Regardless of which interpretation you prefer, a probability must satisfy the three axioms of probability (Kolmogorov, 1933), which are the building blocks of all probability theory.

Definition. The three probability axioms

1. $P(A) \geq 0$ (non-negativity)
2. $P(S) = 1$ (unit measure)
3. $P(A \cup B) = P(A) + P(B)$ if $A \cap B = \emptyset$ (additivity)

define a probability measure that makes it possible to calculate the probability of events.

Note that the probability axioms should be interpreted as follows:

- The first axiom states that the probability of an event $A \subseteq S$ must be non-negative.
- The second axiom states that (a) the probability of an event $A \subseteq S$ must not exceed one, and (b) the probability that at least one elementary event s in the sample space S occurs must equal one. This axiom is a requirement on the sample space S, such that some valid outcome must be observed when the simple experiment is conducted.
- The third axiom states that the probability of mutually exclusive events must be the summation of the probabilities of the events.

Together, these three axioms are all that is needed to compute probabilities for any simple experiment—which is pretty remarkable! For each of the three examples (i.e., coin flipping, card drawing, and dice rolling), you can verify that

(i) the probability of observing any event is greater than or equal to zero

(ii) the probability of observing the entire sample space is equal to one

(iii) the probability of observing mutually exclusive events is the summation of probabilities

Note that it is okay if these points seem somewhat opaque given that we have yet to formally specify the concept of a probability distribution, which we will do in the next section.

4 Probability Distributions

Definition. A probability distribution $F(\cdot)$ is a mathematical function that assigns probabilities to outcomes of a simple experiment. Thus, a probability distribution is a function from the sample space S to the interval $[0, 1]$, which can be denoted as $F : S \to [0, 1]$.

***Example* 13.** Consider the coin flipping example with $n = 3$ coin flips. The sample space is $S = \{0, 1, 2, 3\}$. If we assume that the coin is "fair" (i.e., equal chance of observing Heads and Tails) and that the n flips are "independent" (i.e., unrelated to one another), then the probability of each elementary event is as follows:

s	$P(\{s\})$	Observed flip sequence
0	1/8	(T, T, T)
1	3/8	(H, T, T), (T, H, T), (T, T, H)
2	3/8	(H, H, T), (H, T, H), (T, H, H)
3	1/8	(H, H, H)

Although there are only four elements in the sample space, i.e., $|S| = 4$, there are a total of $2^n = 8$ possible sequences that we could observe when flipping two coins. Given our assumptions, each of the 8 possible sequences is equally likely. As a result, to compute the probability of each $s \in S$, we simply need to count all of the relevant sequences and divide by the total number of possible sequences, which is displayed in the $P(\{s\})$ column of the table. The probability distribution is specified by $P(\{s\})$, such that $P(\{s\})$ defines the probability of observing each elementary event $s \in S$. Note that the probability distribution satisfies the three probability axioms, given that (i) $P(A) > 0$ for any event $A \subseteq S$, (ii) $\sum_{s=0}^{3} P(\{s\}) = 1$, and (iii) $P(\{s\} \cup \{s'\}) = P(\{s\}) + P(\{s'\})$ for any $s, s' \in S$ (with $s \neq s'$).

Some example probability calculations:

- $P(\{0\} \cap \{3\}) = 0$
- $P(\{0\} \cup \{3\}) = P(\{0\}) + P(\{3\}) = 2/8$
- $P(\{a \mid a \text{ is less than } 2\}) = P(\{0\}) + P(\{1\}) = 4/8$
- $P(\{a \mid a \text{ is less than or equal to } 2\}) = \sum_{s=0}^{2} P(\{s\}) = 1 - P(\{3\}) = 7/8$

***Example* 14.** Consider the dice rolling example. The sample space is $S = \{2, 3, \ldots, 11, 12\}$. If we assume that the dice are "fair" (i.e., equal chance of observing each outcome $\{1, \ldots, 6\}$ on a single roll) and that the two rolls are "independent" (i.e., unrelated to one another), then the probability of each elementary event is as follows:

s	$P(\{s\})$	Observed roll sequence
2	1/36	$(1,1)$
3	2/36	$(1,2)$, $(2,1)$
4	3/36	$(1,3)$, $(2,2)$, $(3,1)$
5	4/36	$(1,4)$, $(2,3)$, $(3,2)$, $(4,1)$
6	5/36	$(1,5)$, $(2,4)$, $(3,3)$, $(4,2)$, $(5,1)$
7	6/36	$(1,6)$, $(2,5)$, $(3,4)$, $(4,3)$, $(5,2)$, $(6,1)$
8	5/36	$(2,6)$, $(3,5)$, $(4,4)$, $(5,3)$, $(6,2)$
9	4/36	$(3,6)$, $(4,5)$, $(5,4)$, $(6,3)$
10	3/36	$(4,6)$, $(5,5)$, $(6,4)$
11	2/36	$(5,6)$, $(6,5)$
12	1/36	$(6,6)$

Although there are only 11 elements in the sample space, i.e., $|S| = 11$, there are a total of $6^2 = 36$ possible sequences that we could observe when rolling two dice. Given our assumptions, each of the 36 possible sequences is equally likely. As a result, to compute the probability of each $s \in S$, we simply need to count all of the relevant sequences and divide by the total number of possible sequences, which is displayed in the $P(\{s\})$ column of the table. The probability distribution is specified by $P(\{s\})$, such that $P(\{s\})$ defines the probability of observing each elementary event $s \in S$. Note that the probability distribution satisfies the three probability axioms, given that (i) $P(A) > 0$ for any event $A \subseteq S$, (ii) $\sum_{s=2}^{12} P(\{s\}) = 1$, and (iii) $P(\{s\} \cup \{s'\}) = P(\{s\}) + P(\{s'\})$ for any $s, s' \in S$ (with $s \neq s'$).

Some example probability calculations:

- $P(\{2\} \cap \{12\}) = 0$
- $P(\{2\} \cup \{12\}) = P(\{2\}) + P(\{12\}) = 2/36$
- $P(\{7\} \cup \{11\}) = P(\{7\}) + P(\{11\}) = 8/36$

5 Joint Events

Thus far, we have considered simple experiments where the outcome of interest is a singular event (e.g., the sum of two dice). In such cases, the sample space consists of sample points that are one-dimensional elements. We could easily extend the ideas of probability theory to experiments where the sample points are d-dimensional elements with $d \geq 2$.

Definition. A joint event refers to an outcome of a simple experiment where the sample point is two-dimensional. In this case, the sample points have the form $s = (a, b)$, where a and b are the two events that combine to form the joint event.

***Example* 15.** Suppose that you flip a coin $n = 2$ times and record the outcome of each coin flip (instead of recording the number of heads). In this case, the sample space is $S = \{(a, b) \mid a \in \{H, T\}, b \in \{H, T\}\}$, where a and b denote the outcomes of the first and second coin flip, respectively. Note that the sample space has size $|S| = 4$ and the elementary events are defined as $S = \{(T, T), (H, T), (T, H), (H, H)\}$.

***Example* 16.** Suppose that you pick a card at random from a standard deck of 52 playing cards and record both the value and suit of the card separately. In this case, the sample space is $S = \{(a, b) \mid a \in \{2, 3, \ldots, 9, 10, J, Q, K, A\}, b \in \{\text{Club, Diamond, Heart, Spade}\}\}$. Note that the sample space has size $|S| = 52$, given that a could take 13 different values and b could take 4 different values (and $13 \times 4 = 52$).

***Example* 17.** Suppose that we roll two dice and record the value of each dice (instead of summing the values). In this case, the sample space is $S = \{(a, b) \mid 1 \leq a \leq 6, 1 \leq b \leq 6\}$, where a and b denote the outcomes of the first and second dice roll, respectively. Note that the sample space has size $|S| = 36$. See Example 14 for the 36 elementary events.

Definition. Two events are independent of one another if the probability of the joint event is the product of the probabilities of the separate events, i.e., if $P(A \cap B) = P(A)P(B)$.

Definition. The conditional probability of A given B, denoted as $P(A|B)$, is the probability that A and B occur given that B has occurred, i.e., $P(A|B) = P(A \cap B)/P(B)$.

If A and B are independent of one another, then $P(A|B) = P(A)$ and $P(B|A) = P(B)$. In other words, when A and B are independent, knowing that one of the events has occurred tells us nothing about the likelihood of the other event occurring.

***Example* 18.** For the coin flipping example, if we assume that the coin is fair and the two flips are independent, then $P(s) = (1/2)(1/2) = 1/4$ for any $s \in S$. In other words, if we independently flip a fair coin two times, each of the possible outcomes in the sample space $S = \{(T,T),(H,T),(T,H),(H,H)\}$ is equally likely to occur. Furthermore, if we define $A = \{\text{first flip is heads}\}$ and $B = \{\text{second flip is heads}\}$, then $P(B|A) = P(B) = 1/2$. Thus, the events A and B are independent of one another—which we already knew because we assumed that the two coin flips were independent. Now suppose that we define another event as $C = \{\text{both flips are heads}\}$. Then we have the following probabilities:

- $P(A \cap C) = P(B \cap C) = 1/4$
- $P(A^c \cap C) = P(B^c \cap C) = 0$
- $P(C|A) = P(C|B) = (1/4)/(1/2) = 1/2$
- $P(C|A^c) = P(C|B^c) = 0/(1/2) = 0$

***Example* 19.** For the card drawing example, note that $P(s) = 1/52$ for any $s \in S$, given that we have equal probability of drawing any card in the deck. Suppose that we define the events $A = \{\text{the card is a King}\}$ and $B = \{\text{the card is a face card}\}$. Note that $P(A) = 4/52$ given that there are four Kings in a deck, and $P(B) = 12/52$ given that there are 12 face cards in a deck. The probability of the joint event is $P(A \cap B) = 4/52$ given that $A \subset B$. This implies that $P(A|B) = (4/52)/(12/52) = 4/12$, i.e., if we draw a face card, then the probability of it being a King is 1/3. The opposite conditional probability is $P(B|A) = (4/52)/(4/52) = 1$, i.e., if we draw a King, then it must be a face card. Thus, the events A and B are dependent.

***Example* 20.** For the dice rolling example, if we assume that the dice are fair and the two rolls are independent, then $P(s) = (1/6)(1/6) = 1/36$ for any $s \in S$. Suppose that we define the events $A = \{\text{the sum of the dice is equal to 7}\}$ and $B = \{\text{the first dice is a 1 or 2}\}$. The probabilities of the marginal events are $P(A) = 6/36$ and $P(B) = 2/6$, and the probability of the joint event is $P(A \cap B) = 2/36$ (see Example 14). This implies that $P(A|B) = (2/36)/(2/6) = 2/12$, i.e., if the first roll is 1 or 2, then the probability of the sum being 7 is equal to 1/6. The opposite conditional probability is $P(B|A) = (2/36)/(6/36) = 2/6$, i.e., if the sum of the dice is 7, then the probability of the first roll being 1 or 2 is equal to 1/3. Thus, the events A and B are dependent.

6 Bayes' Theorem

Bayes' theorem (due to Reverend Thomas Bayes, 1763) states that

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \qquad \text{and} \qquad P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

which is due to the fact that $P(A \cap B) = P(B|A)P(A) = P(A|B)P(B)$. Note that Bayes' theorem has important consequences because it allows us to derive unknown conditional probabilities from known quantities. This theorem is the foundation of Bayesian statistics, where the goal is to derive the posterior distribution $P(A|B)$ given the assumed distribution for the data given the parameters $P(B|A)$ and the prior distribution $P(A)$.

7 Basic Probability Properties

Some general results of probability theory:

1. $0 \leq P(A) \leq 1$
2. $P(A^c) = 1 - P(A)$
3. $P(A \cup A^c) = 1$
4. $P(S) = 1$
5. $P(\emptyset) = 1 - P(S) = 0$
6. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
7. $P(A \cup B) \leq P(A) + P(B)$
8. $P(A \cap B) \leq P(A \cup B)$
9. If $A \subseteq B$, then $P(A) \leq P(B)$
10. If $A \subseteq B$, then $P(B \backslash A) = P(B) - P(A)$
11. $P(A|B) = P(A \cap B)/P(B) = P(B|A)P(A)/P(B)$
12. $P(A|B) = P(A)P(B)$ if A and B are independent