

# Introduction to Probability Theory and Statistics

## Probability

Probability theory provides a mathematical foundation to concepts such as “probability”, “information”, “belief”, “uncertainty”, “confidence”, “randomness”, “variability”, “chance” and “risk”. Probability theory is important to empirical scientists because it gives them a rational framework to make inferences and test hypotheses based on uncertain empirical data. Probability theory is also useful to engineers building systems that have to operate intelligently in an uncertain world. For example, some of the most successful approaches in machine perception (e.g., automatic speech recognition, computer vision) and artificial intelligence are based on probabilistic models. Moreover probability theory is also proving very valuable as a theoretical framework for scientists trying to understand how the brain works. Many computational neuroscientists think of the brain as a probabilistic computer built with unreliable components, i.e., neurons, and use probability theory as a guiding framework to understand the principles of computation used by the brain. Consider the following examples:

- You need to decide whether a coin is loaded (i.e., whether it tends to favor one side over the other when tossed). You toss the coin 6 times and in all cases you get “Tails”. Would you say that the coin is loaded?
- You are trying to figure out whether newborn babies can distinguish green from red. To do so you present two colored cards (one green, one red) to 6 newborn babies. You make sure that the 2 cards have equal overall luminance so that they are indistinguishable if recorded by a black and white camera. The 6 babies are randomly divided into two groups. The first group gets the red card on the left visual field, and the second group on the right visual field. You find that all 6 babies look longer to the red card than the green card. Would you say that babies can distinguish red from green?



❖ A pregnancy test has a 99 % validity (i.e., 99 of 100 pregnant women test positive) and 95 % specificity (i.e., 95 out of 100 non pregnant women test negative). A woman believes she has a 10 % chance of being pregnant. She takes the test and tests positive. How should she combine her prior beliefs with the results of the test?

❖ You need to design a system that detects a sinusoidal tone of 1000Hz in the presence of white noise. How should design the system to solve this task optimally?

❖ How should the photo receptors in the human retina be interconnected to maximize information transmission to the brain?

While these tasks appear different from each other, they all share a common problem: The need to combine different sources of uncertain information to make rational decisions. Probability theory provides a very powerful mathematical framework to do so. Before we go into mathematical aspects of probability theory I shall tell you that there are deep philosophical issues behind the very notion of probability. In practice there are three major interpretations of probability, commonly called the frequentist, the Bayesian or subjectivist, and the axiomatic or mathematical interpretation.

### 1. Probability as a relative frequency

This approach interprets the probability of an event as the proportion of times such one event is expected to happen in the long run. Formally, the probability of an event  $E$  would be the limit of the relative frequency of occurrence of that event as the number of observations grows large

$$P(E) = \lim_{n \rightarrow \infty} \frac{n_E}{n} \quad (1.1)$$

where  $n_E$  is the number of times the event is observed out of a total of  $n$  independent experiments. For example, we say that the probability of “heads” when tossing a coin is 0.5. By that we mean that if we toss a coin many many times and compute the relative frequency of “heads” we expect for that relative frequency to approach 0.5 as we increase the number of tosses. This notion of probability is appealing because it seems objective and ties our work to the observation of physical events. One



difficulty with the approach is that in practice we can never perform an experiment an infinite number of times. Note also that this approach is behaviorist, in the sense that it defines probability in terms of the observable behavior of physical systems. The approach fails to capture the idea of probability as internal knowledge of cognitive systems.

## **2. Probability as uncertain knowledge.**

This notion of probability is at work when we say things like “I will probably get an A in this class”. By this we mean something like “Based on what I know about myself and about this class, I would not be very surprised if I get an A. However, I would not bet my life on it, since there are a multitude of factors which are difficult to predict and that could make it impossible for me to get an A”. This notion of probability is “cognitive” and does not need to be directly grounded on empirical frequencies. For example, I can say things like “I will probably die poor” even though I will not be able to repeat my life many times and count the number of lives in which I die poor.

This notion of probability is very useful in the field of machine intelligence.

In order for machines to operate in natural environments they need knowledge systems capable of handling the uncertainty of the world. Probability theory provides an ideal way to do so. Probabilists that are willing to represent internal knowledge using probability theory are called “Bayesian”, since Bayes is recognized as the first mathematician to do so.

## **3. Probability as a mathematical model.**

Modern mathematicians avoid the frequentist vs. Bayesian controversy by treating probability as a mathematical object. The role of mathematics here is to make sure probability theory is rigorously defined and traceable to first principles. From this point of view it is up to the users of probability theory to apply it to whatever they see fit. Some may want to apply it to describe limits of relative frequencies. Some may want to apply it to describe subjective notions of uncertainty, or to build better computers. This is not necessarily of concern to the mathematician.



The application of probability theory to those domains will be ultimately judged by its usefulness.

### 1.1 Intuitive Set Theory

We need a few notions from set theory before we jump into probability theory. In doing so we will use intuitive or “naive” definitions. This intuitive approach provides good mnemonics and is sufficient for our purposes but soon runs into problems for more advanced applications. For a more rigorous definition of set theoretical concepts and an explanation of the limitations of the intuitive approach you may want to take a look at the Appendix.

❖ **Set:** A set is a collection of elements. Sets are commonly represented using curly brackets containing a collection of elements separated by commas.

For example

$$\mathbf{A} = \{1, 2, 3\} \quad (1.2)$$

tells us that **A** is a set whose elements are the first 3 natural numbers. Sets can also be represented using a rule that identifies the elements of the set.

The prototypical notation is as follows

$$\{\mathbf{x} : \mathbf{x} \text{ follows a rule}\} \quad (1.3)$$

For example,

$$\{\mathbf{x} : \mathbf{x} \text{ is a natural number and } \mathbf{x} \text{ is smaller than } 4\} \quad (1.4)$$

❖ **Outcome Space:** The outcome space is a set whose elements are all the possible basic outcomes of an experiment.<sup>1</sup> The sample space is also called **sample space**, **reference set**, and **universal set** and it is commonly represented with the capital Greek letter “omega”, **Ω**. We call the elements of the sample space “outcomes” and represent them symbolically with the small Greek letter “omega”, **ω**.

#### Example 1:

If we roll a die, the outcome space could be

$$\mathbf{\Omega} = \{1, 2, 3, 4, 5, 6\} \quad (1.5)$$

In this case the symbol **ω** could be used to represent either 1,2,3,4,5 or 6.

#### Example 2:

If we toss a coin twice, we can observe 1 of 4 outcomes:



(Heads, Heads), (Heads, Tails), (Tails, Heads), (Tails, Tails). In this case we could use the following outcome space

$$\Omega = \{(H,H), (H, T), (T,H), (T, T)\} \quad (1.6)$$

and the symbol  $\omega$  could be used to represent either (H,H), or (H, T), or (T,H), or (T, T). Note how in this case each basic outcome contains 2 elements. If we toss a coin  $n$  times each basic outcome  $\omega$  would contain  $n$  elements.

❖ **Singletons:** A singleton is a set with a single element. For example the set  $\{4\}$  is a singleton, since it only has one element. On the other hand 4 is not a singleton since it is an element not a set.

❖ **Element inclusion:** We use the symbol  $\in$  to represent element inclusion.

The expression  $\omega \in A$  tells us that  $\omega$  is an element of the set  $A$ . The expression  $\omega \notin A$  tells us that  $\omega$  is **not** an element of the set  $A$ .

**For example,**  $1 \in \{1, 2\}$  is true since 1 is an element of the set  $\{1, 2\}$ . The expression  $\{1\} \in \{\{1\}, 2\}$  is also true since the singleton  $\{1\}$  is an element of the set  $\{\{1\}, 2\}$ .

The expression  $\{1\} \in \{1, 2\}$  is also true, since the set  $\{1\}$  is not an element of the set  $\{1, 2\}$ .

❖ **Set inclusion:** We say that the set  $A$  is included in the set  $B$  or is a **subset** of  $B$  if all the elements of  $A$  are also elements of  $B$ . We represent set inclusion with the symbol  $\subset$ . The expression  $A \subset B$  tells us that both  $A$  and  $B$  are sets and that all the elements of  $A$  are also elements of  $B$ .

For example the expression  $\{1\} \subset \{1, 2\}$  is true since all the elements of the set  $\{1\}$  are in the set  $\{1, 2\}$ . On the other hand  $1 \subset \{1, 2\}$  is not true since 1 is an element, not a set.

❖ **Set equality:** Two sets  $A$  and  $B$  are equal if all elements of  $A$  belong to  $B$  and all elements of  $B$  belong to  $A$ . In other words, if  $A \subset B$  and  $B \subset A$ .

For example the sets  $\{1, 2, 3\}$  and  $\{3, 1, 1, 2, 1\}$  are equal.

❖ **Set Operations:** There are 3 basic set operations:

**1. Union:** The union of two sets  $A$  and  $B$  is another set that includes all elements of  $A$  and all elements of  $B$ . We represent the union operator



with this symbol  $\cup$ .

For example, if  $A = \{1, 3, 5\}$  and  $B = \{2, 3, 4\}$ , then  $A \cup B = \{1, 2, 3, 4, 5\}$ .

More generally

$$A \cup B = \{ \omega : \omega \in A \text{ or } \omega \in B \} \quad (1.7)$$

In other words, the set  $A \cup B$  is the set of elements with the property that they either belong to the set  $A$  or to the set  $B$ .

**2. Intersection:** The intersection of two sets  $A$  and  $B$  is another set  $C$  such that all elements in  $C$  belong to  $A$  and to  $B$ . The intersection operator is symbolized as  $\cap$ . If  $A = \{1, 3, 5\}$  and  $B = \{2, 3, 4\}$  then

$A \cap B = \{3\}$ . More generally

$$A \cap B = \{ \omega : \omega \in A \text{ and } \omega \in B \} \quad (1.8)$$

**3. Complementation:** The complement of a set  $A$  with respect to a reference set  $\Omega$  is the set of all elements of  $\Omega$  which do not belong to  $A$ .

The complement of  $A$  is represented as  $A^c$ . For example, if the universal set is  $\{1, 2, 3, 4, 5, 6\}$  then the complement of  $\{1, 3, 5\}$  is  $\{2, 4, 6\}$ .

More generally

$$A^c = \{ \omega : \omega \in \Omega \text{ and } \omega \notin A \} \quad (1.9)$$

❖ **Empty set:** The empty set is a set with no elements. We represent the null set with the symbol  $\emptyset$ . Note  $\Omega^c = \emptyset$ ,  $\emptyset^c = \Omega$ , and for any set  $A$

$$A \cup \emptyset = A \quad (1.10)$$

$$A \cap \emptyset = \emptyset \quad (1.11)$$

❖ **Disjoint sets:** Two sets are disjoint if they have no elements in common, i.e., their intersection is the empty set. For example, the sets  $\{1, 2\}$  and  $\{1\}$  are not disjoint since they have an element in common.

❖ **Collections:** A collection of sets is a set of sets, i.e., a set whose elements are sets. For example, if  $A$  and  $B$  are the sets defined above, the set  $\{A, B\}$  is a collection of sets.

❖ **Power set:** The power set of a set  $A$  is the a collection of all possible sets of  $A$ . We represent it as  $2^A$ . For example, if  $A = \{1, 2, 3\}$  then

$$2^A = \{ \emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, A \} \quad (1.12)$$



Note that 1 is not an element of  $(A)$  but  $\{1\}$  is. This is because 1 is an element of  $A$ , not a set of  $A$ .

**Collections closed under set operations:** A collection of sets is closed under set operations if any set operation on the sets in the collection results in another set which still is in the collection. If  $A = \{1, 3, 5\}$  and  $B = \{2, 3, 4\}$ , the collection  $C = \{A, B\}$  is not closed because the set  $A \cap B = \{3\}$  does not belong to the collection. The collection  $C = \{\Omega, \emptyset\}$  is closed under set operations, all set operations on elements of  $C$  produce another set that belongs to  $C$ . The power set of a set is always a closed collection.

❖ **Sigma algebra:** A sigma algebra is a collection of sets which is closed when set operations are applied to its members a countable number of times. The power set of a set is always a sigma algebra.

❖ **Natural numbers:** We use the symbol  $\mathbb{N}$  to represent the natural numbers, i.e.,  $\{1, 2, 3, \dots\}$ . One important property of the natural numbers is that if  $x \in \mathbb{N}$  then  $x + 1 \in \mathbb{N}$ .

❖ **Integers:** We use the symbol  $\mathbb{Z}$  to represent the set of integers, i.e.,  $\{\dots, -3, -2, -1, 0, 1, 2, 3, \dots\}$ . Note that if  $x \in \mathbb{Z}$  then  $x + 1 \in \mathbb{Z}$  and  $x - 1 \in \mathbb{Z}$ .

❖ **Real numbers:** We use the symbol  $\mathbb{R}$  to represent the real numbers, i.e., numbers that may have an infinite number of decimals. For example, 1, 2.35,  $-4/123$ , 2, and  $\pi$ , are real numbers. Note that if  $x \in \mathbb{R}$  then  $x + 1 \in \mathbb{R}$  and  $x - 1 \in \mathbb{R}$ .

#### ❖ Cardinality of sets:

- We say that a set is **finite** if it can be put in one-to-one correspondence with a set of the form  $\{1, 2, \dots, n\}$ , where  $n$  is a fixed natural number.
- We say that a set is **infinite countable** if it can be put in one-to-one correspondence with the natural numbers.
- We say that a set is **countable** if it is either finite or infinite countable.
- We say that a set is **infinite uncountable** if it has a subset that can be put in one-to-one correspondence with the natural numbers, but the set itself cannot be put in



such a correspondence. This includes sets that can be put in one-to-one correspondence with the real numbers.

## 1.2 Events

We have defined outcomes as the elements of a reference set  $\Omega$ . In practice we are interested in assigning probability values not only to outcomes but also to sets of outcomes. For example we may want to know the probability of getting an even number when rolling a die. In other words, we want the probability of the set  $\{2, 4, 6\}$ . In probability theory set of outcomes to which we can assign probabilities are called **events**. The collection of all events is called the **event space** and is commonly represented with the letter  $F$ . Not all collections of sets qualify as event spaces. To be an event space, the collection of sets has to be a sigma algebra (i.e., it has to be closed under set operations). Here is an example:

**Example:** Consider the sample space  $\Omega = \{1, 2, 3, 4, 5, 6\}$ . Is the collection of sets  $\{\{1, 2, 3\}, \{4, 5, 6\}\}$  a valid event space?

**Answer:** No, it is not a valid event space because the union of  $\{1, 2, 3\}$  and  $\{4, 5, 6\}$  is the set  $\Omega = \{1, 2, 3, 4, 5, 6\}$  which does not belong to  $F$ . On the other hand the set  $\{\emptyset, \{1, 2, 3\}, \{4, 5, 6\}, \Omega\}$  is a valid event space. Any set operation using the sets in  $F$  results into another set which is in  $F$ .

**Note:** The outcome space  $\Omega$  and the event space  $F$  are different sets. For example if the outcome space were  $\Omega = \{H, T\}$  a valid event space would be  $F = \{\Omega, \emptyset, \{H\}, \{T\}\}$ . Note that  $\Omega = F$ . The outcome space contains the basic outcomes of an experiments. The event space contains sets of outcomes.

## 1.3 Probability measures

When we say that the probability of rolling an even number is 0.5, we can think of this as an assignment of a number (i.e., 0.5) to a set ,i.e., to the set  $\{2, 4, 6\}$ . Mathematicians think of probabilities as function that “measures” sets, thus the name **probability measure**. For example, if the probability of rolling an even number on a die is 0.5, we would say that the probability measure of the set  $\{2, 4, 6\}$  is 0.5. Probability measures are commonly represented with the letter  $P$  (capitalized).





Probability measures have to follow three constraints, which are known as Kolmogorov's axioms:

1. The probability measure of events has to be larger or equal to zero:  $P(A) \geq 0$  for all  $A \in \mathcal{F}$ .

2. The probability measure of the reference set is 1

$$P(\Omega) = 1 \quad (1.13)$$

3. If the sets  $A_1, A_2, \dots \in \mathcal{F}$  are disjoint then

$$P(A_1 \cup A_2 \cup \dots) = P(A_1) + P(A_2) + \dots \quad (1.14)$$

**Example 1: A fair coin.** We can construct a probability space to describe the behavior of a coin. The outcome space consists of 2 elements, representing heads and tails  $\Omega = \{H, T\}$ . Since  $\Omega$  is finite, we can use as the event space the set of all sets in  $\Omega$ , also known as the power set of  $\Omega$ . In our case,  $\mathcal{F} = \{\{H\}, \{T\}, \{H, T\}, \emptyset\}$ . Note  $\mathcal{F}$  is closed under set operations so we can use it as an event space.

The probability measure  $P$  in this case is totally defined if we simply say  $P(\{H\}) = 0.5$ . The outcome of  $P$  for all the other elements of  $\mathcal{F}$  can be inferred:

we already know  $P(\{H\}) = 0.5$  and  $P(\{H, T\}) = 1.0$ . Note the sets  $\{H\}$  and  $\{T\}$  are disjoint, moreover  $\{H\} \cup \{T\} = \Omega$ , thus using the probability axioms

$$P(\{H, T\}) = 1 = P(\{H\}) + P(\{T\}) = 0.5 + P(\{T\}) \quad (1.15)$$

from which it follows  $P(\{T\}) = 0.5$ . Finally we note that  $\Omega$  and  $\emptyset$  are disjoint and their union is  $\Omega$ , using the probability axioms it follows that

$$1 = P(\Omega) = P(\Omega \setminus \emptyset) = P(\Omega) + P(\emptyset) \quad (1.16)$$

Thus  $P(\emptyset) = 0$ . Note  $P$  qualifies as a probability measure: for each element of  $\mathcal{F}$  it assigns a real number and the assignment is consistent with the three axiom of probability.

**Example 2: A fair die.** In this case the outcome space is  $\Omega = \{1, 2, 3, 4, 5, 6\}$ , the event space is the power set of  $\Omega$ , the set of all sets of  $\Omega$ ,  $\mathcal{F} = 2^\Omega$ , and  $P(\{i\}) = 1/6$ , for  $i = 1, \dots, 6$ . I will refer to this as the fair die probability space.

**Example 3: A loaded die.** We can model the behavior of a loaded die by assigning



non negative weight values to each side of the die. Let  $w_i$  represent the weight of side  $i$ . In this case the outcome space is  $\Omega = \{1, 2, 3, 4, 5, 6\}$ , the event space is the power set of  $\Omega$ , the set of all sets of  $\Omega$ ,  $F = 2^{\Omega}$ , and

$$P(\{i\}) = w_i / (w_1 + \cdots + w_6), \quad (1.17)$$

Note that if all weight values are equal, this probability space is the same as the probability space in Example 2.

