MEASURE THEORY & PROBABILITY

The more general foundation of Probability Theory is Measure Theory, which also plays a fundamental role in Real and Complex Analysis. We’ll visit this perspective on Probability Theory throughout the course whenever it provides useful insights on the underlying theory.

BASIC DEFINITIONS

A probability space comprises three parts \((S, \mathcal{F}, P)\):

1. \(S\) is the sample space, the set of all outcomes. (Some texts use \(\Omega\) instead of \(S\)). Ex: For a coin toss experiment, \(S = \{H, T\}\).

2. \(\mathcal{F}\) is the \(\sigma\)-algebra associated with \(S\). It is the set of all subsets of \(S\) (i.e., the set of all events, includes \(S\) and \(\emptyset\)), and is closed under countable unions, countable intersections and complementation. Furthermore, \(\mathcal{F}\) satisfies:
   
   (a) if \(A \in \mathcal{F}\) then \(A^c \in \mathcal{F}\), and
   
   (b) if \(A_1, A_2, \ldots\) are in \(\mathcal{F}\), then their union \(\bigcup_{i=1}^{\infty} A_i\) is also in \(\mathcal{F}\).

Note: Together these conditions imply closure under countable intersections.

3. \(P\) is our probability function \(P : \mathcal{F} \rightarrow [0,1]\). It associates each event (i.e., each subset of \(S\) included in \(\mathcal{F}\)) with a number between 0 and 1. Furthermore, we require that
   
   (a) \(P\) is non-negative \((P(A) \geq P(\emptyset) = 0, \ \forall A \in \mathcal{F})\),
   
   (b) \(P\) is countably additive, i.e., for for a countable, disjoint set of events \(A_1, A_2, \ldots\) then \(P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)\), and
   
   (c) \(P(S) = 1\).

A probability space is a special case of the more general measureable space \((S, \mathcal{F}, \mu)\) which is the conceptual foundation for integration theory taught in standard upper division Real and Complex Analysis courses. Probability functions are a kind of measure, and probability spaces are special only in the sense that \(\mu(S) = 1\). Measure theory deals with the more general case where \(\mu(S) \geq \infty\) (for example, our usual notion of “distance” on \(\mathbb{R}\) implies a standard measure known as Lebesgue Measure, e.g., \(\mu([a,b]) = b - a\), where \(\mu(S) = \mu(\mathbb{R}) = \infty\)). Accordingly, probability spaces are relatively “nice” spaces to work with from an analysis perspective!

NOTE: It’s worth noting these deeper connections to Real and Complex Analysis, as you may encounter some of them in future classes. Additionally, it’s worth remembering that probability functions can be though of as functions that associate probabilities (values in \([0,1]\)) with sets of possible outcomes (i.e., events), and that there are a few different important spaces (e.g., \(S\) vs \(\mathcal{F}\)) we need to keep track of as we progress through the course.

Additional reading on next 3 pages.
Definition 1.2.1. A collection of subsets of $S$ is called a *sigma algebra* (or *Borel field*), denoted by $\mathcal{B}$, if it satisfies the following three properties:

1. $\emptyset \in \mathcal{B}$ (the empty set is an element of $\mathcal{B}$).
2. If $A_1, A_2, \ldots \in \mathcal{B}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{B}$ (the collection is closed under countable intersections).
3. If $A_1, A_2, \ldots \in \mathcal{B}$, then $\bigcap_{i=1}^{\infty} A_i \in \mathcal{B}$ (the collection is closed under countable unions).

The empty set, $\emptyset$, is a subset of any set. Thus $\emptyset \subseteq S$. Property (a) states that the empty set is in $\mathcal{B}$, and property (b) implies that $\mathcal{B}$ is closed under countable intersections. Property (c) also implies that $\mathcal{B}$ is closed under countable unions.

Examples of sigma algebras include the collection of all subsets of $S$, the collection of all open sets of a topological space, and the collection of all measurable sets of a measure space.

Theorem 1.2.2. Let $\mathcal{S} = \{S_1, S_2, \ldots, S_n\}$ be a finite set. Let $\mathcal{B}$ be any sigma algebra of subsets of $\mathcal{S}$. Let $\mathcal{P}(\mathcal{S})$ be the power set of $\mathcal{S}$.

Then $\mathcal{B}$ contains all sets of the form $\bigcup_{i=1}^{n} S_i$ for all $n \in \mathbb{N}$ and $\bigcap_{i=1}^{n} S_i$ for all $n \in \mathbb{N}$.

Proof: By the definition of a sigma algebra, if $S_1, S_2, \ldots, S_n \in \mathcal{B}$, then $\bigcup_{i=1}^{n} S_i \in \mathcal{B}$ and $\bigcap_{i=1}^{n} S_i \in \mathcal{B}$. Therefore, $\mathcal{B}$ contains all finite unions and intersections of sets in $\mathcal{S}$.

Example 1.2.3 (Sigma algebra of intervals). Let $S = (a, b)$, where $a$ and $b$ are real numbers. Then $\mathcal{B}$ contains all intervals $[a, b)$, $(a, b)$, $[a, b]$, $(a, b]$.

Example 1.2.4. Given a sample space $S$ and an associated sigma algebra $\mathcal{B}$, a probability function $P$ is a function $P: \mathcal{B} \rightarrow [0, 1]$ that satisfies:

1. $P(\emptyset) = 0$ for all $\emptyset \in \mathcal{B}$.
2. If $A_1, A_2, \ldots \in \mathcal{B}$ are pairwise disjoint, then $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$.
3. $P(S) = 1$.

The collection of all such probability functions $P$ is the set of probability measures on $\mathcal{B}$.

Example 1.2.5 (Defining probabilities). Consider the simple experiment of tossing a coin, and let $S = \{H, T\}$ be the sample space. Then $P(H) + P(T) = 1$. If we define $P(H) = 1/2$, then $P(T) = 1/2$.

Note that $P(H)$ and $P(T)$ are probabilities of the events $H$ and $T$, respectively. The sum $P(H) + P(T)$ is the probability of the event $S = \{H, T\}$.

Simultaneously solving (1.2.2) and (1.2.3) for $P(H)$ and $P(T)$, we obtain $P(H) = 1/2$ and $P(T) = 1/2$. Therefore, $P(S) = 1$.

In general, if $S$ is an infinite set, it is not an easy task to describe $\mathcal{B}$. However, if $S$ is finite or countable, then $\mathcal{B}$ contains all sets of interest.
(The sum over an empty set is defined to be 0.) Then \( P \) is a probability function on \( \mathcal{B} \). This remains true if \( S = \{s_1, s_2, \ldots\} \) is a countable set.

**Proof:** We will give the proof for finite \( S \). For any \( A \in \mathcal{B} \), \( P(A) = \sum_{i \in S} p_i \geq 0 \), because every \( p_i \geq 0 \). Thus, Axiom 1 is true. Now,

\[
P(S) = \sum_{i \in S} p_i = \sum_{i=1}^{n} p_i = 1.
\]

Thus, Axiom 2 is true. Let \( A_1, \ldots, A_k \) denote pairwise disjoint events. (\( \mathcal{B} \) contains only a finite number of sets, so we need consider only finite disjoint unions.) Then,

\[
P \left( \bigcup_{i=1}^{k} A_i \right) = \sum_{j \in \mathcal{S}} p_j = \sum_{i=1}^{k} \sum_{j \in A_i} p_j = \sum_{i=1}^{k} P(A_i).
\]

The first and third equalities are true by the definition of \( P(A) \). The disjointedness of the \( A_i \)s ensures that the second equality is true, because the same \( p_j \)s appear exactly once on each side of the equality. Thus, Axiom 3 is true and Kolmogorov's Axioms are satisfied.

The physical reality of the experiment might dictate the probability assignment, as the next example illustrates.

**Example 1.2.7 (Defining probabilities–II)** The game of darts is played by throwing a dart at a board and receiving a score corresponding to the number assigned to the region in which the dart lands. For a novice player, it seems reasonable to assume that the probability of the dart hitting a particular region is proportional to the area of the region. Thus, a bigger region has a higher probability of being hit.

Referring to Figure 1.2.1, we see that the dart board has radius \( r \) and the distance between rings is \( r/5 \). If we make the assumption that the board is always hit (see Exercise 1.7 for a variation on this), then we have

\[
P(\text{scoring } i \text{ points}) = \frac{\text{Area of region } i}{\text{Area of dart board}}.
\]

For example

\[
P(\text{scoring 1 point}) = \frac{\pi r^2 - \pi (4r/5)^2}{\pi r^2} = 1 - \left(\frac{4}{5}\right)^2.
\]

It is easy to derive the general formula, and we find that

\[
P(\text{scoring } i \text{ points}) = \frac{(6-i)^2 - (5-i)^2}{5^2}, \quad i = 1, \ldots, 5,
\]

independent of \( \pi \) and \( r \). The sum of the areas of the disjoint regions equals the area of the dart board. Thus, the probabilities that have been assigned to the five outcomes sum to \( 1 \), and, by Theorem 1.2.6, this is a probability function (see Exercise 1.8).

Before we leave the axiomatic development of probability, there is one further point to consider. Axiom 3 of Definition 1.2.4, which is commonly known as the Axiom of Countable Additivity, is not universally accepted among statisticians. Indeed, it can be argued that axioms should be simple, self-evident statements. Comparing Axiom 3 to the other axioms, which are simple and self-evident, may lead us to doubt whether it is reasonable to assume the truth of Axiom 3.

The Axiom of Countable Additivity is rejected by a school of statisticians led by de Finetti (1972), who choose to replace this axiom with the Axiom of Finite Additivity.

**Axiom of Finite Additivity:** If \( A \in \mathcal{B} \) and \( B \in \mathcal{B} \) are disjoint, then

\[
P(A \cup B) = P(A) + P(B).
\]

While this axiom may not be entirely self-evident, it is certainly simpler than the Axiom of Countable Additivity (and is implied by it – see Exercise 1.12).

Assuming only finite additivity, while perhaps more plausible, can lead to unexpected complications in statistical theory – complications that, at this level, do not necessarily enhance understanding of the subject. We therefore proceed under the assumption that the Axiom of Countable Additivity holds.

### 1.2.2 The Calculus of Probabilities

From the Axioms of Probability we can build up many properties of the probability function, properties that are quite helpful in the calculation of more complicated probabilities. Some of these manipulations will be discussed in detail in this section; others will be left as exercises.

We start with some (fairly self-evident) properties of the probability function when applied to a single event.
Section 1.2

BASICS OF PROBABILITY THEORY

A Venn diagram will show why (1.2.7) holds, although a formal proof is not difficult (see Exercise 1.2). Using (1.2.7) and the fact that $A$ and $B \cap A^c$ are disjoint (since $A$ and $A^c$ are), we have

\begin{equation}
(1.2.8) \quad P(A \cup B) = P(A) + P(B \cap A^c) = P(A) + P(B) - P(A \cap B)
\end{equation}

from (a).

If $A \subset B$, then $A \cap B = A$. Therefore, using (a) we have

\begin{equation}
0 \leq P(B \cap A^c) = P(B) - P(A),
\end{equation}

establishing (c).

Formula (b) of Theorem 1.2.9 gives a useful inequality for the probability of an intersection. Since $P(A \cup B) \leq 1$, we have from (1.2.8), after some rearranging,

\begin{equation}
(1.2.9) \quad P(A \cap B) \geq P(A) + P(B) - 1.
\end{equation}

This inequality is a special case of what is known as Bonferroni’s Inequality (Miller 1981 is a good reference). Bonferroni’s Inequality allows us to bound the probability of a simultaneous event (the intersection) in terms of the probabilities of the individual events.

Example 1.2.10 (Bonferroni’s Inequality) Bonferroni’s Inequality is particularly useful when it is difficult (or even impossible) to calculate the intersection probability, but some idea of the size of this probability is desired. Suppose $A$ and $B$ are two events and each has probability .95. Then the probability that both will occur is bounded below by

\[ P(A \cap B) \geq P(A) + P(B) - 1 = .95 + .95 - 1 = .90. \]

Note that unless the probabilities of the individual events are sufficiently large, the Bonferroni bound is a useless (but correct!) negative number.

We close this section with a theorem that gives some useful results for dealing with a collection of sets.

Theorem 1.2.11 If $P$ is a probability function, then

a. \[ P(A) = \sum_{i=1}^{\infty} P(A \cap C_i) \text{ for any partition } C_1, C_2, \ldots; \]

b. \[ P(\bigcup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} P(A_i) \text{ for any sets } A_1, A_2, \ldots. \] (Boole’s Inequality)

Proof: Since $C_1, C_2, \ldots$ form a partition, we have that $C_i \cap C_j = \emptyset$ for all $i \neq j$, and $S = \bigcup_{i=1}^{\infty} C_i$. Hence,

\[ A = A \cap S = A \cap \left( \bigcup_{i=1}^{\infty} C_i \right) = \bigcup_{i=1}^{\infty} (A \cap C_i), \]