Probability Theory Handbook

Rohan Sikand

December 2021

This document contains a collection of important facts, definitions, theorems, examples, etc. about probability theory. It is intended to be short and to the point so that one can use it as a convenient reference[1] (and to see the big picture between the results). Disclaimer: this document provides a collection of items which were not invented by the author and were collected from various sources.

Contents

[1] Indeed, part of the motivation for putting together this collection is so that I can refer to it later on may I forget some of the finer details.

1 Counting

Before we even begin to discuss probability, we need to understanding the basics behind the theory of counting (combinatorics) and set theory.

Definition 1. (**Set notation** [2]) Let A and B be sets.

- The complement[2] of A is $A^C = A' = \bar{A} = \Omega \backslash A$. In words: say $A \subseteq \Omega$. Then A^C is everything in Ω that is not in A.
- "A or B" is the set $A \cup B$.
- "A and B" is the set $A \cap B$.
- A and B are mutually exclusive or disjoint if $A \cap B = \emptyset$.
- If $A \subseteq B$, then A occurring implies B occurring.

[2]Don't worry about the Ω in this definition for now, we will define that later.

2 Probability Spaces

Probability theory is the analysis of random phenomenon. You can think of the theory as a generalization of logical reasoning for randomness. This allows us to reason about uncertainty. We use mathematically structured sets to reason about real world phenomenon modelled by, what we call, **experiments** (random process). We formalize probability via probability spaces which are structured sets representing these experiments.

Definition 2. (**Probability space** [3, 7]) A probability space, defined by the triple $(\Omega, \mathcal{F}, P)$, is a set with added structure which formally models a random experiment. A probability space consists of three elements:

1. **Sample space,** Ω: the set of all possible outcomes of the experiment. The constituents of the experiment.

2. **Event space,** $\mathcal{F}$: the set of all potential results (events) of the experiment. For discrete distributions, this is often the power set[3] (set of all subsets) of Ω. Events are often represented as capital letters (e.g. $A \in \mathcal{F}$).

3. **Probability function,** P: a probability distribution/function/measure which assigns each event in the event space a probability, which is a number between 0 and 1 ($P : \mathcal{F} \to \mathbb{R}$). That is, for each event $A \in \mathcal{F}$, we associate a number using a function, $P(A)$, that measures the probability (frequentist) or degree of belief (Bayesian) that the event will occur. The result of $P(A)$ is then called the probability of A. We say that an event "occurs" in a trial if the outcome of that trial is an element of the specified event set.

This is best illuminated via the following canonical example.

Example. Say we want to model the real world process of throwing a singular die. We define the sample space as all of the possible outcomes that the die can land on: $\Omega = \{1, 2, 3, 4, 5, 6\}$. We define the event space as the power set of Ω which is the set of all events. Note that there is no construct in the definition which states the size of each subset needs to be the same[4].

[3] Though this isn't always the case. For example, we may want to design an experiment such that certain events are restricted from occurring.

[4] There is something called an elementary event which is an event consisting of exactly one outcome from Ω. The point I am trying to make here is that there can exist events in the event space that are not elementary.

For example, we could have the event $\{6\}$ meaning "the die lands on 6" and we could also have the event $\{2, 4, 6\}$ meaning "the die lands on an even number". Finally, we'd construct a probability function which maps each event in the event space to a probability.

It is important to note the distinction between the sample space and the event space. The sample space contains the *outcomes* of the experiments. The event space contains the all of the *sets* of possible outcomes (arranged in any fashion containing one or multiple elements) of the experiments. An outcome is a possible result of an experiment or trial [7]. When we throw a die, it can only land on *one* of six possible values. So of course, we can assign a probability to each of these six values, but we can also assign a value to any subset of these outcomes (i.e. the die lands on an odd number which would be a set of size 3). Another way to think about events is to think of an event as some subset of the sample space that we ascribe meaning to (e.g. all even numbers) [4]. It is important to be able to distinguish between an outcome and an event[5].

Remark. (frequentist vs. Bayesian) Probability can be viewed from two different perspectives: from the frequentist perspective or the Bayesian perspective. At first thought, the frequentist perspective might seem more familiar and intuitive. But it doesn't encompass all possible phenomenon. The frequentist perspective is that the probability of each event in the event space occurring is the fraction of times that the event will occur if the experiment is repeated many (approaching infinite) times in real life [1]. Coin flipping and die rolling are two common examples. However, this perspective, as mentioned, does not cover experiments that are not repeatable. For example, what if we want to measure the probability that a candidate will win an election or if we want to measure the probability that a patient has a virus? In this case, we take the Bayesian approach meaning we think of probabilities as "beliefs that are updated according to some set of rules as new knowledge is acquired" [1]. For example, we can interpret the probability of the candidate winning the election as as quantification of a personal belief (subjectivity) and/or reasonable expectation based on our current knowledge [6].

How do we know what types of functions can be used for probability functions (for probability spaces), P? What functions are valid probability functions? The requirements/constraints for P are given by the Kolmogorov axioms defined below.

[5]For more explanations, see here and here.

Definition 3. (**Axioms of Probability** [4]) Let $(\Omega, \mathcal{F}, P)$ be a probability space and let A and B be events in the event space $\mathcal{F}$ $(A, B \in \mathcal{F})$. The probability function, P, must satisfy the following axioms which we accept as truth.

- **Axiom 1**:

$$0 \leq \mathrm{P}(A) \leq 1$$

  That is, the resulting probability of an event occurring must be a number between 0 and 1.

- **Axiom 2**:

$$P(\Omega) = 1$$

  That is, the probability that at least one of the elementary events in the sample space will occur is 1 [7]. Logically, this means that an outcome from any trial of the experiment will be in the sample space, Ω meaning that all outcomes must be from the sample space, Ω. Intuitively, this also means that the sum of the probabilities for each element in the sample space is equal to 1.

- **Axiom 3**:

$$\mathrm{P}(A \text{ or } B) = \mathrm{P}(A) + \mathrm{P}(B)$$

  where A and B are mutually exclusive events (they cannot both occur at the same time meaning the sets are disjoint[6]). In words, this means that the probability of A or B occurring is the sum between the probability that A will occur and the probability that B will occur. This generalizes beyond just two events [5]. Specifically, for any sequence of mutually exclusive events $(E_1, E_2, ...)$:

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P\left(E_i\right).$$

From these axioms, we can formulate and prove a couple useful identities.

[6] An example is coin tossing. A trial can result in heads or tails but not both.

Proposition.

$$\mathrm{P}\left(E^{\mathrm{C}}\right) = 1 - \mathrm{P}(E).$$

That is, the probability of event E not happening (defined as the compliment of $P(E)$) is equal to 1 minus the probability of the event happening.

We can prove this directly from the axioms.

Proof.

$$\begin{aligned} \mathrm{P}(S) &= \mathrm{P}\left(E \text{ or } E^{\mathrm{C}}\right) \\ \mathrm{P}(S) &= \mathrm{P}(E) + \mathrm{P}\left(E^{\mathrm{C}}\right) \\ 1 &= \mathrm{P}(E) + \mathrm{P}\left(E^{\mathrm{C}}\right) \\ \mathrm{P}\left(E^{\mathrm{C}}\right) &= 1 - \mathrm{P}(E) \end{aligned}$$

□

Proposition. Let E and F be events. Then,

$$\text{if } E \subseteq F, \text{ then } \mathrm{P}(E) \leq \mathrm{P}(F).$$

That is, if the event E is contained in the event F, then the probability of E is no greater than the probability of F [5][7].

Proof. The key realization is to construct

$$F = E \cup \left(F \cap E^{C}\right)$$

From axiom 3, we obtain

$$P(F) = P(E) + P\left(F \cap E^{C}\right)$$

Finally, since $P\left(F \cap E^{C}\right) \geq 0$, $\mathrm{P}(E) \leq \mathrm{P}(F)$.

□

Brief discussion of what a random variable is and how it differs from probability: Let's say you want to know the result of a random process– not the probability of the random process; just the result. We can store this unknown result in what is called a random variable. This allows us to reason about the random process without it occurring yet. This differs from probability because we don't care about finding the likelihood that this process will occur (it is not an event in the event space). We simply just

[7]This bit was taken directly from Ross [5].

want to reason mathematically about the random process so we assign a variable to it.

For example, say we roll a die three times and want to know the sum of the die rolls. This process is fundamentally random so we can't assign a concrete number to it. For this reason, we can assign a variable (placeholder) to store this unknown value. Note that we aren't asking anything about the probability here–we don't care about the sum being a specific number, we just want to store the unknown number (sum). Then, this will allow us to reason mathematically about this variable (i.e. calculate the probability of the die rolls summing to 7).

Resources

Below, I enumerate some resources for further reading.

- Dexter Chua's lecture notes on Probability (link) [2].
- The canonical text in the U.S.: Sheldon Ross's textbook [5].
- CS 109 Course Reader (link) [4].
- For a more theoretical treatment (i.e. proof-based), see the lecture notes for Stanford's Math 151 course by Sourav Chatterjee [1].

References

[1] Sourav Chatterjee. *Lecture notes for Math 151*. Winter 2020.

[2] Dexter Chua. *Part IA — Probability*. 2015.

[3] Marc Peter Deisenroth, A Aldo Faisal, and Cheng Soon Ong. *Mathematics for machine learning*. Cambridge University Press, 2020.

[4] Chris Piech and Stanford University Department of Computer Science. *Course Reader for CS109*. Sept. 2021.

[5] Sheldon M Ross. *A first course in probability*. Pearson, 2014.

[6] Wikipedia contributors. *Bayesian probability — Wikipedia, The Free Encyclopedia*. `https://en.wikipedia.org/w/index.php?title=Bayesian_probability&oldid=1049806208`. 2021.

[7] Wikipedia contributors. *Probability space — Wikipedia, The Free Encyclopedia*. `https://en.wikipedia.org/w/index.php?title=Probability_space&oldid=1061316054`. 2021.