Statistics and Probability

free and open source book written for educational purposes

Joseph Mehdiyev

A book written for the author's educational purposes

Preface

About the Book

This book is sort of a big notebook to make (or force) the author to self study and understand the field of Probability and Statistics. Concepts and topics are explained with details and examples. Almost all theorems and lemmas have proofs. There are some exceptions on basic or similiar theorems where proof is only a sketch. I have to note that this book is an educational and fun project for the author himself. Through the book, author tries to explain the topics to *himself* .Be careful using the book as the main learning material, since the writer himself is not an expert in the field, there may be mathematical errors in the book.

I like to explain the mathematical concepts in more "traditional" way. I don't like long and complex theorems, lemmas with comically big proofs that reader must pray to understand. Through the book, I try to explain the concepts in everyday language. Of course, rigorous proofs are also provided as they are still an important part of mathematics.

To learn the field and write this book, I used various books from known authors, countless mathematics forums about statistics and probability, and wikipedia (duh-duh) articles. These are some of the books I used majority of time:

- Larry Wasserman - All of Statistics - A Concise Course in Statistical Inference.
- Dimitri Bertsekas And John N Tsitsiklis - Introduction To Probability
- Mathematical Statistics with Applications by Dennis Wackerly, William Mendenhall, and Richard L. Scheaffer
- Joseph K. Blitzstein - Introduction to Probability
- Ross, Sheldon - First Course in Probability

I want to note that I did not, by any means, plagiarize any contents, diagrams or other things. I simply wrote whatever I learnt through the the brainstorm I had. Theorems and proofs may be similar, but I believe it is acceptable since I can't rigorously find another way of defining theorems and proving them.

Book's source

Maybe you may already know this, this book is fully open source with its pictures and tex file shared in author's github. You may use the source code for whatever purposes you want to use it for. If you want to contribute, please send a pull request from the github.Currently the book is in development.

How to use the Book

As the book is precise and short, you may use the book as a revisit or a secondary material. The book shortly and simply explains the concepts and ideas. Important concepts' proofs are provided. However, other proofs explaned in sentences rather than other classic rigorous proofs.

Coding stuff

The statistical images are being generated by **python's matplotlib**, while other sort of diagrams are mix of **latex's tikz** or **matplotlib**. Moreover, there are practical examples with **python** of probability and statistical concepts through the book. You can get more information from the book's github page.

Contents

Part I

Probability

Chapter 1

Introduction to Probability

The concept "probability" is used very often in everyday language to describe the chance of something happening. Mathematically, Probability is a language to quantify uncertainty. This chapter will introduce necessary and basic concepts and namely, **Probability Theory**. We will start the chapter about interpretations of probability.

1.1 Interpretations of Concept of Probability

We will briefly skim through this section.

In a theorical environment i.e tossing a coin with fifty to fifty chance, probabilities can be represented as fractions. This called **Theorical Probability**. However, in practical applications, there are two major categories on interpretations:**Frequency** and **Bayesian**

Frequentist Probability, as name implies, gets its name from frequency. In this perspective probability is interpreted same as frequency. Repeating the experiment high number of times, one may find approximate probability of an *event*. This is the dominant form of probability that is taught in schools and universities.

Bayesian Probability, however, takes its name from **Bayes' Rule**, which we will learn later. In this intersection, the probabilities represents the degree of belief on an event i.e the more information or conditions we have about an event, its probability changes.

There are also other intersections, but they are not that widely used nor useful. It is enough to know above concepts.

1.2 Set Theory

Set Theory is a branch of mathematics that studies *sets*, which we will define shortly. This branch is, like other parts of mathematics, very deep and complex. We will learn only the most important concepts , which is in high-school level, needed to understand later sections and chapters.

We will quickly introduce the concepts and briefly explain them. The reader may skip this section if they already know about sets and their basic properties.

Sets

A **Set** is a collection of different objects, which are called *elements* of the set. The sets are notated as capital letters such as S. If x is an element of a set S, we write $x \in S$. Otherwise we write $x \notin S$. A set with no elements is called **empty set** and is notated as $\emptyset$.

If $x_1, x_2, ..., x_n$ are the elements of the set S, we write:

$$S \in \{x_1, x_2, ..., x_n\}$$

If S is set of all even numbers smaller than 12, we can draw the diagram as:

We can specify our set as a selection from a larger set. If we want to write the set of all even integers, we can write (Here the set of integers is the universal set):

$$S = \{n \in \mathbb{Z} : \frac{n}{2} \text{ is an integer}\}$$

If a set A's elements are also the elements of B, we say that A is a **subset** of B. We can notate it as:

$$A \subseteq B$$

If a set A is subset of B, but is not equal to B, we say that A is **proper subset** of B. We can notate it as:

$$A \subsetneq B$$

Set operations

Union of sets A,B is a set that contains the elements of A and B:

$$A \cup B = \{n : n \in A \vee n \in B\}$$

We can visualize the sets in 2D with circles and their intersections.

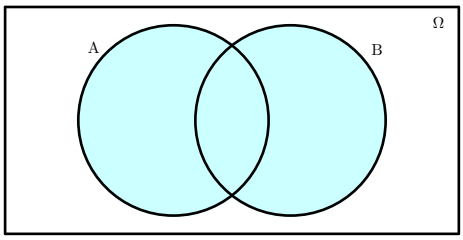

Figure 1.1: $A \cup B$

Intersection of sets A,B is a set that contains both the elements of A and B:

$$A \cap B = \{n : n \in A \wedge n \in B\}$$

Sample Space and Events

The Sample Space, usually denoted as S or Ω, is the *set* of all possible outcomes of an experiment. It is also called **universal set**. Subsets of Ω are called **events**. A sample element of Ω is denoted as ω.

Example 1.2.1. If we toss a six sided dice once, then $\Omega = \{1, 2, 3, 4, 5, 6\}$, the even that the side is even is $A = \{2, 4, 6\}$ while $\omega \in \{1, 2, 3, 4, 5, 6\}$

Example 1.2.2. If we toss a two sided coin twice, then

$$\Omega = \{(HH), (TT), (HT), (TH)\} \ \wedge \ \omega \in \{(HH), (TT), (HT), (TH)\}$$

Example 1.2.3. If we toss a 2 sided coin forever, then

$$\Omega = \{\omega = (\omega_1, \omega_2, ...) : \ \omega_i \in \{H, T\}\}$$

Example 1.2.4. Let E be the event that only even numbers appear in the six sided dice toss. Then,

$$E = \{2, 4, 6\}$$

With the new definition, we can make more set operation: **complement** of the event A is a set of elements Ω that do not belong to A.

$$A^c = \{n : n \in \Omega \wedge n \notin A\}$$

difference of the set A from B is a set of elements of A that do not also belong to B

$$A \setminus B = A \cap B^c$$

we say that $E_1, E_2, ..., E_N$ are **disjoint** if

$$A_i \cap A_j = \emptyset$$

A partition of Ω is a sequence of disjoint events such that

$$\bigcup^{\infty} E_i = \Omega$$

Similar to **monotone functions**, we define **monotone increasing** sequence of sets $A_1, A_2, ...$ as the sequence of sets such that $A_1 \subset A_2 \subset ...$ and $\lim_{n\to\infty} A_n = \bigcup A_i$

Moreover, we can define certain rules similar to the rules of algebra:

Commutative laws	$A \cup B = B \cup A$
Associative laws	$(A \cup B) \cup C = A \cup (B \cup C)$
Distributive laws	$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

And lastly, **DeMorgan's laws** states that

$$\left(\bigcup_{i=1}^{n} A_i\right)^c = \bigcap_{i=1}^{n} A_i^c$$

$$\left(\bigcap_{i=1}^{n} A_i\right)^c = \bigcup_{i=1}^{n} A_i^c$$

Which is, in my opinion, very intuitive and can be easily understood with sketching venn diagrams. These are all of the terminology and notations we will be using for learning the probability.

1.3 Probability Law

To show the probability of a event A, we assign a real number $P(A)$ or $\mathbb{P}(A)$ in some textbooks, called **probability of** A. In other words, $P()$ is a unique function with unique properties that inputs an event A, and outputs its probability.

To qualify as probability, P must satisfy 3 axioms:

Axiom 1 $P(A) \geq 0$ for every A

Axiom 2 $P(\Omega) = 1$

Axiom 3 If $A_1, A_2, ...$ are disjoint:

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

Let's explain the axioms. The first axiom is very simple, a probability can't be negative, since the meaning of the word probability. Second axiom is also very simple, the probability of any possible outcomes happening is 1, since there must be a outcome at the end of the experiment. Third axiom, assume we have 2 disjoint sets. Then

$$P(A \cup B) = P(A) + P(B)$$

This is true simply because sets are disjoint. Similarly, we can use induction to prove the above property for n sets. Proving for infinite sets are out of scope of this section, therefore we will skip it.

We can derive many properties from these axioms. These are the most simple and intuitive ones:

$$\begin{aligned} P(\emptyset) &= 0 \\ A \subset B &\implies P(A) \leq P(B) \\ 0 \leq P(A) &\leq 1 \\ P(A^c) &= 1 - P(A) \end{aligned}$$

And a less obvious property:

Lemma 1.3.1. For events A and B,

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

Proof. We can rewrite $A \cup B$ as union of $A \setminus B$, $B \setminus A$, and $A \cap B$, since these are the slices of the thing we want to begin with. Moreover, these slices are disjoint, therefore we can apply our third axiom (P is additive):

$$\begin{aligned} P(A \cup B) &= P\big((A \setminus B) \cup (B \setminus A) \cup (A \cap B)\big) \\ &= P(A \setminus B) + P(B \setminus A) + P(A \cap B) \\ &= P(A \setminus B) + P(A \cap B) + P(B \setminus A) + P(A \cap B) - P(A \cap B) \\ &= P(A) + P(B) - P(A \cap B) \end{aligned}$$

□

1.4 Discrete Probability Distribution

Discrete Probability distribution is the mathematical description of probability of events, that are subsets of **finite or countable infinite** set Ω. If each outcome is equal, then probability of getting 2 even numbers from tossing a six sided dice, which is $\frac{1}{4}$,is an example of this. We can generalize this for event A of finite Ω,

$$P(A) = \frac{|A|}{|\Omega|}$$

This is the equation almost everybody gets taught in high-school. We can calculate probability of getting heads from tossing a coin, getting a red ball from a box, getting a number from tossing n sided coin and so on. To compute this probability, we first have to count $|\Omega|$ and $|A|$.

For simple experiments, it is rather easy just do count by finger. However, sometimes things get rather complex and we have to use new tools to count them. For example, how many possible outcomes are there from tossing a coin $10^{64^{100}}$ times? We will learn more about counting techniques in Chapter 1.7.

1.5 Independent Events

If we flip a six sided dice twice, probability of getting 2 even numbers is $\frac{1}{4}$, which can be found easily just by counting. However, one may guess that we can find the probability for one dice, then square it, which gets the same answer, $\frac{3}{6} \times \frac{3}{6} = \frac{1}{4}$.

This is a prime example of **Independent Events**. The first roll and the second roll are not depended on each other. Whatever the results in first roll can't influence the result in second roll.

The formal definition of independence is,

Definition 1.5.1. Two events A and B are **independent** if

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

But how can we know the events are *Independent*? Sometimes, it is rather simple, we know it by logic. Probability of the author being successful is not depended on tossing a coin, it is just simple logic.

In almost all cases, simple logic is enough to determine this property. Another property, is that *disjoint events are never Independent*. Other than that, we have to manually check if the events satisfy the above equation.

1.6 Conditional Probability

Conditional Probability, as the name implies, is the probability of an event with a condition. More precisely, **Conditional Probability** is the probability of an event A, given that another event B is already occurred. In such probability, the sample space is reduced to B's, while we want to find probability of A from B's space (Which increases of probability of A, since sample space is also reduced). We can show this neatly in venn diagram:

PICTURE HERE

Here are some examples:

Example 1.6.1. If we tossed a six sided dice one time, and we rolled an even number B, what is the probability of getting number 2, event A?

Since the first toss' result is already happened, we know that $\Omega_{reduced} = \{2, 4, 6\}$ and $A = \{2\}$, then $P(A)_{\Omega_{reduced}} = \frac{1}{3}$.

If there wasn't any condition, the probability of getting 2 would be $\frac{1}{6}$. Simply, in a simple probability we defined a new condition and sort of updated our measurement to $\frac{1}{3}$. This is an important idea in Probability and Statistics, which we will revisit shortly in **Bayes' Rule**

We can show the conditional probability of A given B as:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad \text{for } P(B) \neq 0$$

If we revisit to our simple probability equation, this equation starts making sense since $P(B)$ becomes our reduced sample space, while $P(A \cap B)$ is our event fancily written for condition property.

It is a very common mistake to thing $P(A|B) = P(B|A)$, which is easy to understand why just by looking to either venn diagrams or the equations we defined. Moreover, if A and B are independent from each other, then $P(A|B) = P(A)$, which comes from the definition of independence, B can't effect A's probability.

GENERALIZE THE THING!!!

1.7 Bayes' Theorem

In this section, we will learn about **Bayes' Theorem**, an important concept about probability. This rule is widely used by scientists and programmers. But, what is this rule exactly? Why is it useful?

Bayes' Rule, in simple words, helps to calculate conditional probabilities. It helps us to view probabilities in a degree of belief. I highly recommend watching 3blue1brown's video about this concept (since visual teaching will always be more practical).

We firstly begin by introducing the simple version of the theorem:

Theorem 1.7.1 (Simplified Bayes' Theorem)**.**

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

Proof. We apply the definition of conditional probability twice:

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

Using above properties directly gives our theorem. □

Let's try to comprehend the theorem more practically. The theorem can be understood as "Updating the probability of A with a new condition B". You may think this is an obvious fact and couldn't be that useful. However, let's give some examples that are actually very ambigious without the theorem.

Example 1.7.1. Steve is a middle aged man living in USA and he is very patient and curious. He also likes debate with people. Which is more likely about Steve: A known mathematician that earned a noble prize or a plumber?

Majority of people would immediately answer "the mathematician", however there is a bigger chance he is a plumber. The reason people get wrong on these questions is because they think that these specific attributes directly corresponds to a smart, wise man. However, they also forget that the number of noble prize winner, middle aged mathematician men that lives in USA is quite low (maybe even zero, I don't really know). The attributes may be likely to the mathematician, however there is also a low chance that a plumber can have these specific attributes. Also considering there are almost 300k plumbers, the numbers add up.

To not make these kind of mistakes, we must think these attributes, or events as new updates on our main probability, which is a man either being mathematician or a plumber. That is the core idea of Bayes' Theorem.

When using the Bayes' Theorem, it is not always practical to directly calculate the $P(A)$ or $P(B)$. Therefore we need another tool, called **Law of Total Probability** which states that.

Theorem 1.7.2 (Law of Total Probability)**.** Let $A_1, A_2, ..., A_n$ be partition of Ω. Then for any event B,

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$$

Proof. Let $C_i = A_i \cap B$. Then we know that $C_1, C_2, ..., C_n$ are the partition of B. Therefore using the partition property,

$$P(B) = \sum_{i=1}^{n} P(C_i) = \sum_{i=1}^{n} P(A_i \cap B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$$

Last step is consequence of conditional probability definition of $P(B|A_i)P(A_i) = P(B \cap A_i)$ □

This theorem becomes very handy in practical situations. Moreover, with the help of this theorem we can generalize our Bayes' Theorem,

Theorem 1.7.3 (Bayes' Theorem)**.** Let $A_1, A_2, .., A_n$ be a partition of Ω such that $P(A_i) > 0$. For $P(B) \neq 0$ and for any $i = 1, 2, ..., n$,

$$P(A_i|B) = \frac{P(A_i) \cdot P(B|A_i)}{P(B)} = \frac{P(A_i) \cdot P(B|A_i)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)}$$

Proof. Similar to proof of Theorem 1.7.1, We use definition of conditional probability and lastly apply Theorem 1.7.2 in the last step. □

1.8 Counting