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 Discrete Versus Continuous Concepts

Before we even begin with our concepts. we must learn the difference between the terms **Discrete** and **Continuous** probabilities. Through the book, we will use these terms many times. The Mathematics bluntly can be divided into two distinct categories: *Continuous* mathematics is the study of the objects are uncountable values i.e real numbers, intervals of real numbers and so on; *Discrete* mathematics are study of countable objects. Take probability for example. The probability of simple head and tails experiment is considered discrete, while the probability of weighting 150 kg from intervals 100kg and 200kg is continuous. We will dive deep into these concepts later on, however it is nice to know these terms' meanings beforehand.

1.3 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.

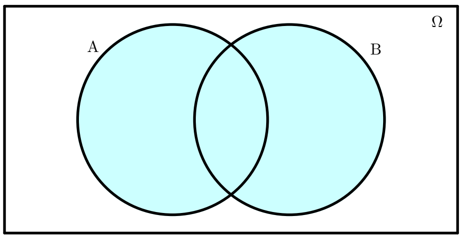


Figure 1.1: $A \cup B$

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

Example 1.3.2. If we toss a two sided coin twice, then

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

Example 1.3.3. If we toss a 2 sided coin forever, then

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

Example 1.3.4. 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.

1.4 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:

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

And a less obvious property:

Lemma 1.4.1. For events A and B,

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

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):

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

□

1.5 Probability Distribution

There are two kinds of Probability Distribution: **Discrete** and **Continous** 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.

Continuous Probability Distribution is similar to its discrete counterpart, however the outcomes are uncountably infinite. Consequently, any probability of selected outcome is 0. Only the events that include these outcomes, making a countable collection of events, have probability themselves. We will learn more about this property in **Chapter 2**.

1.6 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 1.6.1. 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.

1.7 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:

Here are some examples:

Example 1.7.1. 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 think $P(A|B) = P(B|A)$, 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.

1.8 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 1.8.1 (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 1.8.1. 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 1.8.2 (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 $P(B|A_i)P(A_i) = P(B \cap A_i)$ $\square$

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

Theorem 1.8.3 (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. $\square$

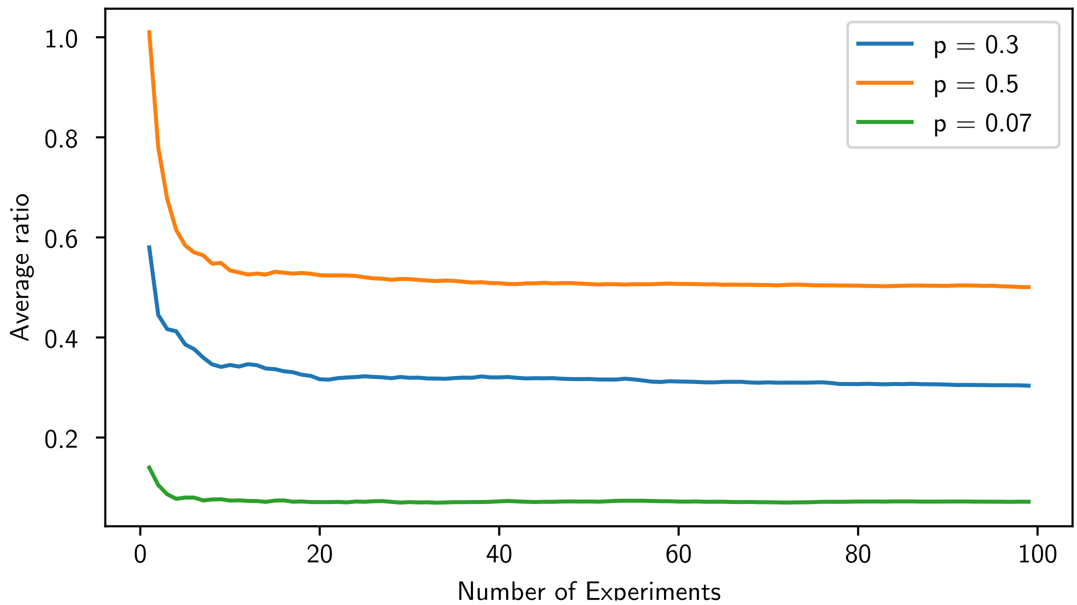
1.9 Exercises

1. Suppose a coin has probability p of getting heads. Logically, if we tossed coin in a large sum many times, and take average of our data, the proportion of heads would be near p. Write a simulation with n tosses and p probability to show the claim.

Solution Source

```
import numpy as np
import matplotlib.pyplot as plt
```

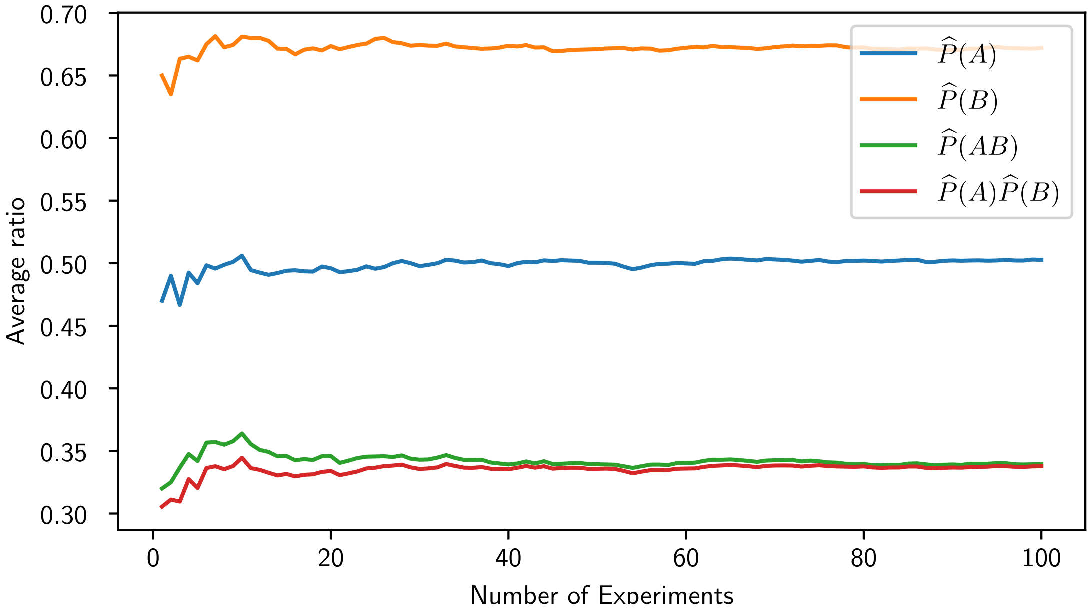
```
# These are our parameters
n = 100
p1 = 0.3
p2 = 0.5
p3 = 0.07
numberOfExperiments = n

# We already discussed about binomial distributions. Coin flipping is one of
these.
# This function does the experiment n times, saves them in result variable as
an array.
result1 = np.random.binomial(n = n,p = p1, size = numberOfExperiments)
result2 = np.random.binomial(n = n,p = p2, size = numberOfExperiments)
result3 = np.random.binomial(n = n,p = p3, size = numberOfExperiments)


#
nn = np.arange(0, n, 1)
plt.figure(figsize=(6.5, 3.5))
plt.plot(nn, np.cumsum(result1) / (nn*n), label='p = 0.3')
plt.plot(nn, np.cumsum(result2) / (nn*n), label='p = 0.5')
plt.plot(nn, np.cumsum(result3) / (nn*n), label='p = 0.07')

plt.legend(loc='upper right')
plt.ylabel("Average ratio")
plt.xlabel("Number of Experiments")
plt.savefig('src/chapter1/fig1.pgf')
```



2. Let $A = \{2, 4, 6\}$ and $B = \{1, 2, 3, 4\}$. Let $\widehat{P}(A)$ be the proportion of times we get A in the experiment. Then by simulation, prove that $\widehat{P}(AB) = \widehat{P}(A)\widehat{P}(B)$.
Solution Source

```
import numpy as np
import matplotlib.pyplot as plt

n = 100
numberOfExperiments = n
```

```
# This is a numpy thing. It will be used for generating probability
distributions.
# Basically makes the probability possible
rng = np.random.default_rng()

# Here we draw random integers in interval (1,6) for n times, and repeat the
process "numberOfExperiment" times
result = rng.integers(low = 1, high = 7, size = (n, numberOfExperiments) )

# This loop counts the number of capA we get in one experiment, and loops
through the total experiments.
totalCapA = []
for element in result:
    # Outputs Bool values depending on the input array, i.e [2,4,6]
    initialCapA = np.isin(element, [2,4,6])
    # Number of Bool values.
    totalCapA.append(initialCapA.sum())

# Repeat the same process for capB and capAB
totalCapB = []
for element in result:
    initialCapB = np.isin(element, [1,2,3,4])
    totalCapB.append(initialCapB.sum())
totalCapAB = []
for element in result:
    initialCapAB = np.isin(element, [2,4])
    totalCapAB.append(initialCapAB.sum())

# A simple array [1,2,...,n]
experimentOrder = np.arange(1,n+1,1)


plt.figure(figsize=(6.5, 3.5))

# The fancy expressions here just takes the average value.
# X axis is "experimentOrder" while the Y axis is our fancy expressions.
plt.plot(experimentOrder, np.cumsum(totalCapA)/(experimentOrder *n), label =
"$\widehat{P}(A)$")
plt.plot(experimentOrder, np.cumsum(totalCapB)/(experimentOrder *n), label =
"$\widehat{P}(B)$")
plt.plot(experimentOrder, np.cumsum(totalCapAB)/(experimentOrder *n), label =
"$\widehat{P}(AB)$")
plt.plot(experimentOrder,
np.multiply(np.cumsum(totalCapA),np.cumsum(totalCapB))/((experimentOrder
*n)*(experimentOrder *n)), label = "$\widehat{P}(A)\widehat{P}(B)$")

plt.ylabel("Average ratio")
plt.xlabel("Number of Experiments")
plt.legend(loc='upper right')
plt.savefig('src/chapter1/fig2.pgf')
```



Graph shows that the red and green graphs will converge to the same value for a larne n.

3. Repeat the above experiment for dependent events.

Chapter 2

Random Variables

From the first chapter, we have been using events and sample spaces to develop the idea of probability and calculating it. But, in practical world, we have to link the events and sample spaces to **data**. This concept is called "Random Variable" or R.V shortly. A Random Variable, in informal terms, places the Ω in real line so we can work with it more easily. There are still events, but in terms on Random Variables now.
From now on, we may write Random Variables as R.V or R.V. or R.Vs shortly for convenience.

2.1 Introduction to Random Variables

A Random Variable describes the data or the outcome ω as a real number. There is a reason this concept exists, since it opens new concepts for practical applications.
Let's begin with the formal definition of *Random Variable.*

Definition 2.1.1. A **Random Variable X** is a function,

$$X \ : \ \Omega \to \mathbb{R}$$

That assigns a real number $X(\omega)$ to each outcome ω.

This concept is heavily used instead of sample spaces . From now on, sample space will be mentioned rarely. Think this way, when we work on functions in algebra or sometimes in calculus, we don't think about about the domain of the function, but the properties of function itself. Here are some examples to understand the concept better.

Example 2.1.1. Flip a fair coin n times. Let X represent the number of heads we get. Then, X is a random variable that takes values $\{0, 1, 2, ..., n\}$.

Example 2.1.2. Toss a fair six sided dice 2 times. Let X be the sum of the two rolls we get. Then, X is a random variable that takes values $\{2, 3, 4, ..., 12\}$.

Example 2.1.3. A students wants to write a real number in intervals $[0, 1]$. Let X be the number student writes. Then, X is also a random variable that takes any real numbers in that interval.

As you may guess, this extremely looks similar to events. Random Variables also have *Independence, Conditional Random Variable, a probability function* and so on. Additionally , Random Variables can be either **Discrete** or **Continous**.

Discrete Random Variable's range is finite or countably infinite. The first two examples we gave are Discrete. Continuous Random Variables's range is uncountably infinite like the third example

I want to emphasize that Random Variables are neither random or a variable, they are functions. It is a bit hard to grasp the idea of this concept, so I highly recommend lurking in mathematical forums and try to understand it (that is what I did). But in short, we use Random Variables instead of outcomes, since Random variables are **numbers**. Numbers are easier to work with, we can process the numbers, do algebraic operations to them, also they have a structure that outcomes do not. Turn **Example 2.1.1** to in sample space and events language, which is easier to work with? Bunch of H, T or just a number?

2.2 Distribution Functions CDF,PMF,PDF

PMF

We define **Cumulative Distribution Function** as,

Definition 2.2.1. The **Cumulative Distribution Function** or shortly **CDF** is a function $F_X : \mathbb{R} \to [0,1]$ such that

$$F_X(x) = P(X \leq x)$$

Remark : Every R.V (discrete and continuous) have CDF. For this reason, we can use CDF for unified treatment of R.V properties (that is, generalized concepts for all R.V).

In informal terms, CDF is the probability that X will take a value less than or equal to x. This property holds both for continuous and discrete R.V.

Later on the book we will learn that CDF practically contains all the information about R.V, including continuous ones. Let's look at a example for CDF.

Example 2.2.1. We toss a fair coin two times. Let X represent the number of heads we get. Then CMF of X is,

$$F_X(x) = \begin{cases} 0 & x < 0 \\ 1/4 & 0 \leq x < 1 \\ 3/4 & 1 \leq < 2 \\ 1 & x \geq 2 \end{cases}$$

The variable x can get **any real numbers**, such as $2, 4.14$ and π. It a bit tricky, they simply take the values from corresponding inequalities.
Now, let's look at some properties of CDF,

Theorem 2.2.1. Let X have *CDF* F and Y have *CDF* G. If $F(x) = G(x)$ for all x, then,

$$P(X \in A) = P(Y \in A) \quad \text{for all } A$$

Theorem 2.2.2. the function $F : \mathbb{R} \to [0,1]$ is a CDF for some R.V if and only if F satisfies three conditions:

1. F is non-decreasing
2. F is normalized i.e
$$\lim_{x \to -\infty} F(x) = 0 \quad \wedge \quad \lim_{x \to \infty} F(x) = 1$$
3. F is right continuous.

CDF and PDF

Similar to probabilities of Events, we can calculate probability of X , depending on discrete or Continuous with functions called **Probability Mass Function** and **Probability Density Function**, shortly **PMF** and **PDF** respectively,

Definition 2.2.2. If X is discrete, and it takes *countably* values $\{x_1, x_2, .., x_n\}$ we define

Probability Mass Function of X as follows:

$$f_X(x) = P(X = x)$$

Remark: $\{X = x\}$ are disjoint events that form partition of Ω.

With the properties of probability, we have $f_X \geq 0$ for all $x \in \mathbb{R}$ and $\sum_i f_X(x_i) = 1$. Let's revisit our Example 2.2.1

Example 2.2.2. We toss a fair coin two times. Let X represent the number of heads we get. Then CMF of X is,

$$f_X(x) = \begin{cases} 1/4 & x = 0 \\ 1/2 & x = 1 \\ 1/4 & x = 2 \\ 0 & \text{otherwise} \end{cases}$$

Moreover, for any set of real numbers, S, we have

$$P(X \in S) = \sum_{x \in S} f_X(x)$$

Since all $\{X = x\}$ are disjoint.

We can apply similar rules to continuous R.Vs,

Definition 2.2.3. If X is continuous, we can represent the probability distribution of X with,

$$P(a < X < b) = \int_a^b f_X(x)dx$$

Function f_X is called **Probability Density Function** or PDF as shortly.

Nothing new here really, we just change the properties of PMF that we can use it on continuous R.Vs. Now, let's look at some examples,

You may noticed that CDF is similar to PMF and PDF. Indeed, the are related, CDF is just sum of these functions we defined over some interval x.

Definition 2.2.4. CDF is related to PMF and PDF. For discrete R.Vs,

$$F_X(x) = P(X \leq x) = \sum_{x_i \leq x} f_X(x_i)$$

And for continuous R.Vs,

$$F_X(x) = \int_{-\infty}^{x} f_X(x)dx$$

And $f_X(x) = F_X^{'}(x)$ for for all differentiable points x.

2.3 Some Important Random Variables and their PMF and PDF

There are some specific examples of R.V. that are very useful in practical applications. We will learn most important ones.

Bernoulli R.V

Consider a cheating coin toss which has of probability p for head, and a probability of $1 - p$ for tails. **Bernoulli R.V** outputs two values: 1 if head and 0 if tails,

$$X = \begin{cases} 1 & \text{if head} \\ 0 & \text{if tails} \end{cases}$$

Then its PMF is,

$$f_X(x) = \begin{cases} p & \text{if } x = 1 \\ 1-p & \text{if } x = 0 \end{cases}$$

This R.V is very simple and easy to understand. However, there are tons of applications in real life (Any True or False situations, for example) and it is very handy to construct more complex R.V.

Binomial R.V

Instead of tossing a cheating coin one time, we toss n times, that is generalized version of **Bernoulli R.V**. With same rules, p is probability of heads and $1-p$ is probability of tails. Let X be the number of heads we get. Then PMF of X is,

$$f_X(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x \in \{0, 1, 2, ..., n\}$$

2.4 Multivariate Distribution

In practical word, we often work with multiple R.V in the same experiment, or the sample space. This can be a medical research with multiple tests, where tests are related with each other with the same sample space Ω and the same probability.
We can apply multivariate CDF as

Definition 2.4.1. For n R.V $\{X_1, X_2, .., X_n\}$, the multivariate CDF $F_{X_1,X_2,...,X_n}$ is given by,

$$F_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n) = P(X_1 \leq x_1, ..., X_n \leq x_n)$$

There is nothing fancy here, actually. We simply redefine CDF in general sense for n R.Vs.

Similarly, we can define multivariate PMF as,

Definition 2.4.2. For n discrete R.Vs of $\{X_1, X_2, X_n\}$, the multivariate PMF $f_{X_1,X_2,...,X_n}$ is given by,

$$f_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n) = P(X_1 = x_1 \ \wedge ... \wedge \ X_n = x_n)$$

The Properties and theorems are similar, but are generalized for n R.Vs.

2.5 Marginal Distribution

If more than one variable is defined in an experiment, it is important to distinguish between the multivariate probability of $(X_1, X_2, .., X_n)$ and individual probability distributions of $X_1, X_2, .., X_n$

Definition 2.5.1. If $(X_1, X_2, ..., X_n)$ are joint distributions with PMF $f_{X_1,X_2,..,X_n}$, then we define marginal distribution as,

$$f_{X_1} = P(X_1 = x_1) = \sum_{x_1 \ constant} P(X_1 = x_1, .., X_n = x_n) = \sum_{x_1 \ constant} f_{X_1,X_2,..,X_n}(x_1, x_2, ..., x_n)$$

Examples

2.6 Independence

Similar to events, R.Vs also can be independent,

Definition 2.6.1. Two R.Vs X and Y are **independent** if, for every A and B,

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$$

To check Independence, we need to check the above question for every subsets A, B. Additionally, we have the theorem,

Theorem 2.6.1. Let X and Y have PMF f_{X_y}. THen X and Y are independent only and only if ,

$$f_{X,Y}(x, y) = f_X(x)f_Y(y)$$

2.7 Conditioning

Similar to events, R.V X can also have conditional distributions given that we have $Y = y$. We show the conditionality with,

Definition 2.7.1. We can show conditional distribution of X respect to Y with,

$$P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}$$

Moreover we can also define **conditional PMF** as,

Definition 2.7.2. PMF of X conditional respect to Y can be written as

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

Examples

2.8 Transformations of a Random Variable

In some applications, we really are interested in distributions of some function of X. We call this concept **Transformation of X**.

Definition 2.8.1. Let X be R.V with PDF/PMF f_X and CDF F_X. Let $Y = r(X)$ i.e $Y = X^2$ or $Y = \ln X$. We call $Y = r(X)$ **transformation of x**.

If Y is discrete, PMF is given by,

$$f_Y(y) = P(Y = y) = P(r(X) = y) = P(\{x\ :\ r(x) = y\}) = P(X \in r^{-1}(y))$$

If Y is continuous, we first calculate CDF and find derivative of it.

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(r(X) < y) \\ &= P(\{x\ :\ r(x) \leq y\}) = P(A_y) \\ &= \int_{A_y} f_X(x)dx \end{aligned}$$

And the last step, $f_Y(y) = F^{'}(y)$.

We can also generalize this concepts for Multivariate distributions, which we just increase dimensions we work with (too lazy, add this later).

Chapter 3

Expectations and Invariance

3.1 Expectation of a Random Variable

The distribution of X contains all the probabilistic data we need about X. However, we need additional tools to describe these data more cleanly.

One of these tools is **Expectation**, or **Expected Value** or **Mean** of X.

Definition 3.1.1. The **expected value** of X is defined as,

$$\mathbb{E}(X) = \begin{cases} \sum xf(x) & \text{if } X \text{ is discrete} \\ \int xf(x)dx & \text{if } X \text{ is continous} \end{cases}$$

If expected value is infinite, we say that expected value of X doesn't exist.
We can also combinte both the notations into a whole generalized equation with a notation,

$$\mathbb{E}(X) = \int xdF(x) = \mu = \mu_X$$

We have discussed that $dF(x) = f(X)$. in the second chapter. **Important Note**. Expectation, by nature, is a theorical mean of the variables we get. It is sometimes possible to get mean that you can't get in a practical settings.

By definition of probability, sum of all $f(x)$ is is simply 1. Then, the above equation is weighted mean of X, which is what we wanted to convey.

Example 3.1.1. Suppose that we have a discrete R.V X to describe the probability of getting heads from tossing a coin 3 times. Let c.d.f of X be f. Then,

$$X = \begin{cases} f(0) = 1/8 \\ f(1) = 3/8 \\ f(2) = 3/8 \\ f(3) = 1/8 \end{cases}$$

Let's use our above formula to calculate $\mathbb{E}(X)$,

$$\mathbb{E}(X) = \frac{1}{8} \cdot 0 + \frac{3}{8} \cdot 1 + \frac{3}{8} \cdot 2 + \frac{1}{8} \cdot 3 = 1.5$$

This number shows that if we repeat our experiment for a very long time, the mean of the heads we got woud be (or approach to) 1.5.
Observe that weighted mean is equivalent to arithmetic mean. Because gettings $X = 2$ is simply getting $X = 1$ two times.

But, what if $Y = g(X)$ and we want to compute $E(Y)$? We have a theorem for that,

Theorem 3.1.1 (Law of the Unconscious Statistician)**.** Let $Y = g(X)$. Then,

$$\mathbb{E}(Y) = \mathbb{E}(g(X)) = \int g(x)dF_X(x)$$

The general proof of this theorem is out of the scope of this book. Comparing this to original expectation equation, we can see that the only thing that changes is $g(x)$ and x, which intuitively makes sense if you think about it. In transformations, probabiltiies remains unchanged, while the result of probabiltiies gets transformed by a function.
Moreover, for a special case $g(x) = I_A(x)$, where $I_A(x) \in \{0, 1\}$ depending on $x \in A$, then,

$$\mathbb{E}(I_A(X)) = \int I_A(x)dF_X(x) = \int_A dF_X(x) = \mathbb{P}(X \in A)$$

This means that probability is special case of expectation, which makes sense, considering probability itself is some average by definition.

Definition 3.1.2. We call n-th **raw moment** of X as

$$\mu_n = \mathbb{E}(X^n) = \int x^n dF_X(x)$$

If $\mathbb{E}(|X^k|)$ is infinite, then k^{th} moment do not exist.
We also define **k-th central moment** as moments about its mean μ i.e $\mathbb{E}[(X - \mu)^k]$. Additionally, **k-th standardized moments** as $\frac{\mathbb{E}[(X-\mu)^n]}{\sigma^n}$.

The 1st moment, the 2nd central moment, 3rd and 4rd standarized moments are called mean (expected value), **variance** ,**skewness** and **kurtosis** in order. We will learn more about them in later chapters.

The moments are very useful and practical. Although there are infinitely many moments, only smaller ones are important for practical purposes. We already know the first moment and its significance.

Properties of Expectation

Theorem 3.1.2 (Non-negativity)**.** If $X \geq 0$ is a R.V, then $\mathbb{E}(X) \geq 0$.

Proof. By definition of expectation, we have

$$\mathbb{E}(X) = \int x dF_X(x) \geq 0$$

since by definition, $dF_X(X) \geq 0$ and $x \geq 0$. □

Theorem 3.1.3 (Linearity)**.** For **any** random variables $X_1, X_2, ..., X_n$ and constants $a_1, a_2, ..., a_n$, we have

$$\mathbb{E}\left[\sum_i^n a_i X_i\right] = \sum_i^n a_i \mathbb{E}(X_i)$$

Proof. We will first prove the theorem for $n = 2$ with $X, Y..$ $n = 1$ is trivial.

$$\begin{aligned}\mathbb{E}\Big[a_1 X_1 + a_2 Y\Big] &= \int (a_1 x + a_2 y) dF_{X,Y}(x,y) \\ &= \int (a_1 x) dF_X(x) + \int (a_2 y) dF_Y(y) \\ &= a_1 \int x dF_X(x) + a_2 \int y dF_X(y) \\ &= a_1 \mathbb{E}(X) + a_2 \mathbb{E}(Y)\end{aligned}$$

The second line is the direct consequence of marginality. With induction, $n \geq 3$ is also true, however I will omit the solution for the sake of briefity. □

This theorem is very useful and very practical.

Theorem 3.1.4 (multiplicity)**.** For **independent** R.V $X_1, X_2, ..., X_n$, we have

$$\mathbb{E}\left(\prod_{i=1}^{n} X_i\right) = \prod_{i=1}^{n} \mathbb{E}(X_i)$$

Proof. Similiar to last one, we will use indiction. $n = 1$ is trivial. For $n = 2$, let R.V be X, Y. Remember that independence has property $dF_{X,Y}(x,y) = dF_X(x) \cdot dF_Y(y)$.

$$\mathbb{E}(XY) = \int (xy) dF_{X,Y}(x,y) = \int xy dF_X(x) dF_Y(y) = \int y dF_Y(y) \int x dF_X(x) = \mathbb{E}(X)\mathbb{E}(Y)$$

For the sake of briefity, I won't show the induction part. □

3.2 Conditional Expectation

Suppose that we we want to calculate mean of X when $Y = y$. This is called conditional expectation, similar to conditional R.V and probability.

Definition 3.2.1. conditional expectation of X by $Y = y$ is given by,

$$\mathbb{E}(X|Y = y) = \int x dF_{X|Y}(x|y)$$

Note that $\mathbb{E}(X|Y)$ is a R.V itself since we don't know value of Y beforehand, or more precisely Y is a "function".

Theorem 3.2.1 (Law of total Expectations)**.** for all R.V X and Y,

$$\mathbb{E}[\mathbb{E}(Y|X)] = \mathbb{E}(Y) \qquad \text{and} \qquad \mathbb{E}[\mathbb{E}(X|Y)] = \mathbb{E}(X)$$

Proof. It is direct consequence of definition of conditional expectation and the fact that $dF(x,y) = dF(x)dF(y|x)$

$$\mathbb{E}[\mathbb{E}(Y|X)] = \int \mathbb{E}(Y|X = x) dF(x)$$

writelater □

3.3 Variance

We have dicussed about the expectation, a way of showing a property of a distribution. However, the expectation alone doesn't convey much. We have another tool called '**Variance**'. Variance, in layman terms, describes how value of random variable varies are spread in the graph. Or in other terms, the distance between the expectation value.

We can define variance ass,

Definition 3.3.1. Let X be a R.V with mean $\mu = \mathbb{E}(X)$. The Variance of X, denoted as $\mathbb{V}(X)$ or $\text{Var}(X)$ or σ^2 is the 2^{nd} central moment and is defined by,

$$\sigma^2 = \mathbb{E}[(X-\mu)^2] = \int (x-\mu)^2 dF(x)$$

We also define **standart deviation** as $\text{sd}(X) = \sqrt{\sigma^2} = \sigma$.

The standart deviation and variance convey the same information. They both represent the spread of our data. The difference between them is purely mathematical. The variance is more useful in mathematical applications, where standart deviation is very intuitive and practical. mathisfun explains it very well.

Calculating variance directly can be complicated and tedious directly sometimes. We can derive a theorem from the original definition for practical purposes.

Theorem 3.3.1. Let X be a random variable. Then,

$$\sigma^2 = E([(X-\mu)^2]) = E(X^2) - \mu^2$$

Proof. It is derived directly by algebraic manipulation and basic calculus,

$$\begin{aligned}
\sigma^2 &= \int (x-\mu)^2 dF(x) \\
&= \int x^2 dF(x) - 2\mu \int x dF(x) + \mu^2 \int dF(x) \\
&= \int x^2 dF(x) - \mu^2 \\
&= \mathbb{E}(X^2) - \mu^2
\end{aligned}$$

□

3.4 Conditional Variance

Definition 3.4.1. Let $\mu = \mathbb{E}(X|Y=y)$. The **conditional variance** is defined as,

$$\mathbb{V}(X|Y=y) = \int (x-\mu)^2 dF_{X|Y}(x|y)$$

The conditional variance tells us how much of spread is left after We use $Y = y$. Reminder that $\mathbb{V}(X|Y)$ is a R.V itself since Y is a sort of "function" here.

Theorem 3.4.1 (Law of Total Variance)**.** for any R.V X, Y, it is always true that,

$$\mathbb{V}(Y) = \mathbb{E}\big[\mathbb{V}(Y|X)\big] + \mathbb{V}\big(\mathbb{E}[Y|X]\big)$$

We have stated before that $V(Y|X)$ and $E(Y|X)$ are random variables, not numbers. Therefore We compute variance and expectation of these random variables, and add them up to get the variance $V(Y)$.

3.5 Covariance and Corelation

Ley X and Y be R.V. **Covariance** and **Corelation** describes the linear relationship between X and Y.

Definition 3.5.1. If X and Y are R.V with mean μ_X, μ_Y and standart deviations σ_X, σ_Y, we define **covariance** as,

$$\text{Cov}(X, Y) = \mathbb{E}\Big((X - \mu_X)(Y - \mu_Y)\Big)$$

and **corelation** as,

$$\rho = \rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

Notice that $\text{Cov}(X, X) = \mathbb{V}(X)$ and $\rho_{X,X} = 1$.

Similiar to variance, calculatig covariance can be tedious. We can derive a better formula by simple algebraic manipulations,

Theorem 3.5.1. For all random variables with non-infinite means, we have

$$\text{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

Proof. Similar to Variance one, we have,

$$\begin{aligned}
\text{Cov(X, Y)} &= \mathbb{E}\Big((X - \mu_X)(Y - \mu_Y)\Big) \\
&= \mathbb{E}(XY - X\mu_Y - Y\mu_X + \mu_X\mu_Y) \\
&= \mathbb{E}(XY) - \mu_Y\mathbb{E}(X) - \mu_X\mathbb{E}(Y) + \mu_X\mu_Y \\
&= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)
\end{aligned}$$

□

Theorem 3.5.2. For all random variables with non-infinite means, we have,

$$-1 \leq \rho_{X,Y} \leq 1$$

Proof. It is direct consequence of Cauchy-Schwarz inequality. □

Chapter 4

Inequalities

not yet sure, may delete this chapter

Chapter 5

Convergence of Random Variables

Will write after,

Part II

Statistics