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.

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.

1.1 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.

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

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

For simplicity we also write $A \cap B = AB$.

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 1.1.1. 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 1.1.2. If we toss a two sided coin twice, then

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

Example 1.1.3. If we toss a 2 sided coin forever, then

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

Example 1.1.4. 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.

1.2 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 1.2.1. 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):

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

□

1.3 Probability Distributions

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 use counting techniques, namely combinatorics. However, the book assumes the reader has knowledge of Combinatorics, therefore we won't introduce the concept here.

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.

After we learn about **Random Variables**, we will talk about specific distributions.

1.4 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 1.4.1. 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.

Example 1.4.1. Let $A = \{2, 4, 6\}$, $B = \{1, 2, 3, 4\}$. Since $P(A)P(B) = P(AB)$, they are independent.

Example 1.4.2. Let $A = 2, 4, 6$, $B = \{2, 4, 5\}$. Since $P(A)P(B) \neq P(AB)$, they are dependent.

Note that even though $A \cap B \neq \emptyset$, in above examples, the result is not the same. The independence merely shows that another event can't change other event's probability, even though intuitively it makes no sense.

1.5 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 1.5.1. 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.

1.6 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 1.6.1 (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 1.6.1. 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 1.6.2 (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 1.6.3 (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 $i = 1, 2, ..., n$,

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

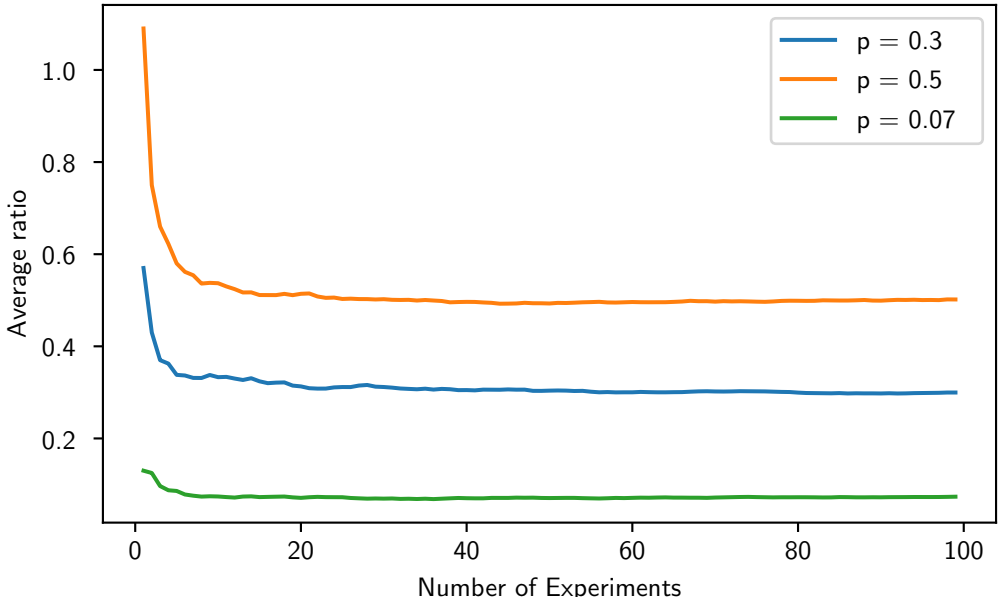
Proof. Similar to proof of Theorem 1.7.1, We use definition of conditional probability and lastly apply Theorem 1.7.2 in the last step. □

1.7 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, 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')
```

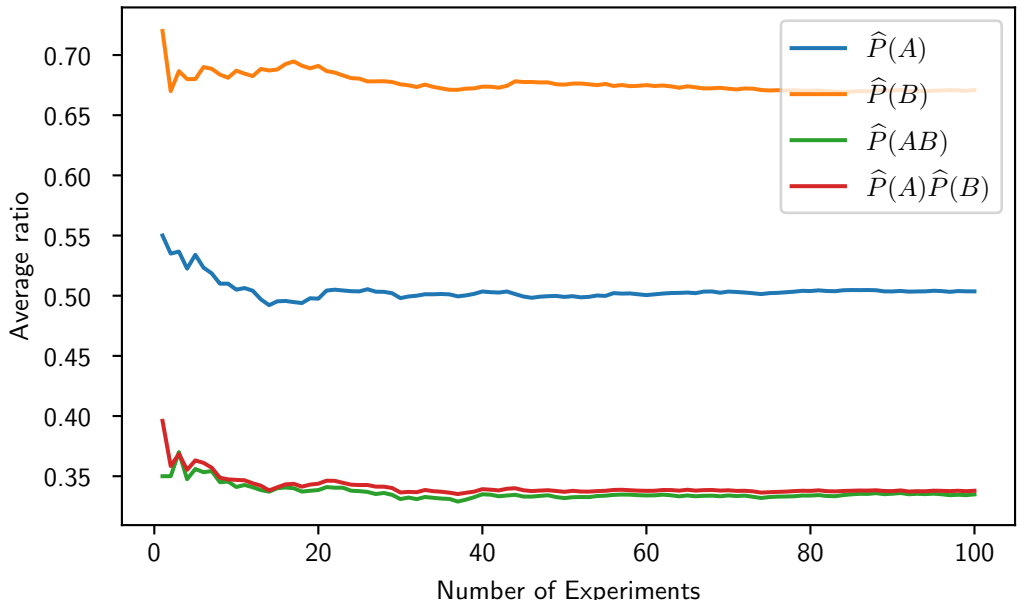


2. Let $A = \{2, 4, 6\}$ and $B = \{1, 2, 3, 4\}$. Let $\widehat{P}(A)$ be the proportion of times we get A in the experiment. Then by simulation, prove that $\widehat{P}(AB) = \widehat{P}(A)\widehat{P}(B)$.

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, 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 large n. Note that the result will be the same even though A and B are different events, even if they are dependent.

Chapter 2

Random Variables

"A random variable can be compared with the holy roman empire: The Holy Roman Empire was not holy, it was not roman, and it was not an empire." - Some random guy in internet

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 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.
In **layman terms**, a random variable is a way to assign a numerical code to each possible outcome. A r.v is **neither random, nor a variable.** It is just a function.

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 coin. We know that $\Omega = \{H, T\}$. A r.v X might assign $X(H) = 1$ and $X(T) = 0$. That is, heads is "coded" as 1 and tails is "coded" as 0.

Example 2.1.2. 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 2.1.3. 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 2.1.4. 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 c.d.f, p.m.f, p.d.f, p.p.f

c.d.f and p.p.f

We define **Cumulative Distribution Function** as,

Definition 2.2.1. The **Cumulative Distribution Function** or shortly **c.d.f** 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 c.d.f. For this reason, we can use c.d.f for unified treatment of r.v properties (that is, generalized concepts for all r.v). Moreover, c.d.f contains all the information about r.v, both continous and discrete ones. That is why c.d.f is very useful, even in practical world.

In informal terms, c.d.f 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.

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

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

The variable x can get **any real numbers**, such as $2, 4.14$ and π. It is just that in discrete case the probability equals to 0 (Which in other words converts discrete to continous function?). It a bit tricky, they simply take the values from corresponding inequalities.
Now, let's look at some properties of c.d.f,

Theorem 2.2.1. Let X have *c.d.f* F and Y have *c.d.f* 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 c.d.f 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.

Remarks. The first and second properties are simple and intuitive, therefore we will ignore them (I know it is not the mathematical way, but whatever).

Third property, however, is worth having discussion about. This property directly follows from the inequality $\leq$. We could even define c.d.f with strict inequality $F = P(X < x)$, and it would still work. It is matter of convention. □

Definition 2.2.2. Quantile Percent point function, or shortly p.p.f, is defined as inverse of c.d.f i.e,
$$Q(x) = F^{-1}(x)$$

c.d.f and p.d.f

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 **p.m.f** and **PDF** respectively,

Definition 2.2.3. If X is discrete, and it takes *countably* values $\{x_1, x_2, .., x_n\}$ we define **Probability Mass Function** of X as follows:
$$f_X(x) = P(X = x)$$
Note that $P(X = x)$ is a function, not a number. We have to specify x first to get a number.
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 c.d.f 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 2.2.4. 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 p.m.f that we can use it on continuous r.vs. Now, let's look at some examples,

You may noticed that c.d.f is similar to p.m.f and PDF. Indeed, the are related, c.d.f is just sum of these functions we defined over some interval x.

Definition 2.2.5. c.d.f is related to p.m.f 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.

Note that this definition is heavily used instead of direct definitions above, since we can work with c.d.f only, and derive it to get needed functions.

2.3 Important Random Variables and their distribution

Definition 2.3.1. If X has distribution A, we write

$$X \sim A$$

Usually A depends on some fixed numbers to define properly, we call them **parameters**. For example, the distribution **Bernoulli**(p) has parameter p. We show parameters in c.d.f and p.m.f as,

$$f(x; parameters) \quad \text{and} \quad F(x; parameters)$$

There are some specific examples of r.v. that are very useful in practical applications. We will show most important ones, and briefly discuss them. In later chapters, we will learn more about them. Note that we will write the notation with the name of the distribution.

Degenerate distribution or Point mass distribution: $X \sim \delta_a$

Consider tossing coin or dice where all the sides show the same value. The p.m.f is ,

$$f_X(x; \delta_a) = 1 \quad \text{for } x = a$$

You might guess why it is called degenerate sometimes. It is not random, but the distribution satisfies the definitions!

Discrete Uniform distribution

This distribution is the one of the most known ones. When there are finitely many values and each of them have the same probability, then $p = \frac{1}{n}$. Simple coin tossing, dice rolling are prime example of these. The p.m.f is,

$$f_X(x) = \frac{1}{n}$$

Where $x \in \{1, 2, ..., n\}$. Nothing new here. for other cases, $f_X(x) = 0$.

Bernoulli distribution: $X \sim \text{Bernoulli}(p)$

This distribution describes "Yes or No" type of experiments such as coin flipping. Therefore, $P(X = 1) = p$, $P(X = 0) = 1 - p$. We can also calculate p.m.f,

$$f_X(x; p) = p^x(1 - p)^{1-x} \quad \text{for } x \in \{0, 1\}$$

Binomial distribution: $X \sim \text{Binomial}(n, p)$

This distribution is generalized form of **Bernoulli distribution**. Similar to Bernoulli, this distribution describes "Yes or no" type of experiments, but for n times of tries e.g tossing a coin n times. Assuming tries are independent of each other, we can show p.m.f as,

$$f_X(x; n, p) = \binom{n}{x} p^x(1 - p)^{n-x} \quad \text{for } x \in \{0, 1, ..., n\}$$

Notice that Binomial$(1, p)$ = Bernoulli(p).

Geometric distribution: $X \sim \text{Geom}(p)$

This distribution is also specified with Bernoulli. The geometric distribution describes the probability of the first occurrence of success requires after x independent trials e.g getting the first head after x tosses. The p.m.f is,

$$f_X(x;p) = (1-p)^{x-1}p \quad \text{for } x \geq 1$$

Poisson distribution: $X \sim \text{Poisson}(\lambda)$

This distribution is mainly used for counts of events like photons hitting a detector in a time interval, number of car accidents, students achieving a low and high mark on exam, or number of pieces of chewing gum on a tile of a sidewalk. its p.m.f is,

$$f_X(x;\lambda) = e^{-\lambda}\frac{\lambda^x}{x!} \quad \text{for } x \geq 0$$

Usually, $\lambda = rt$ where r is average rate the events occur and t is the time interval. The r.v X represents the number of events.

Unfiorm distribution: $X \sim \text{Uniform}(a,b)$

The p.d.f of X is defined as,

$$f_X(x;a,b) = \frac{1}{b-a} \quad \text{for } x \in [a,b]$$

Normal (Gaussian) distribution: $X \sim N(\mu,\sigma^2)$ **or** $X \sim \mathcal{N}(\mu,\sigma^2)$

This distribution is one of the most popular ones even for non-mathematician, layman people. The famous IQ graph, badly made "memes" are example of this. This distribution plays important role in statistics and probability. Moreover, we can observe this distribution in nature. p.d.f is defined as,

$$f_X(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{for } x \in \text{R}$$

We will learn about μ and σ in later chapters. For now, just assume they are some random parameters.
We define **standart Normal distribution** as $N(0,1)$, r.v as Z. This specific distribution is very important, so much that we show its p.d.f and c.d.f with new notation, namely $\phi(z)$ and $\Phi(z)$. There is no closed form expression for $\Phi(z)$. In modern days, the programming libaries calculate them.
It can be shown that we can show **any normal probabilities** we want with $\Phi(z)$.

Exponential distribution: $X \sim \text{Exp}(\lambda)$

This distribution is continous analogue of the geometric distribution. This distribution (and the ones after this) has complex properties so we will be brief with this. We define its c.d.f as ,

$$f_X(x;\lambda) = \lambda e^{-x\lambda}$$

Gamma distribution: $X \sim \text{Gamma}(\alpha,\beta)$

First, we start with a definition,

Definition 2.3.2. For $\alpha > 0$, we define **Gamma function** as,

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}dt$$

It is generalized form of more simple and specific version,

$$\Gamma(n) = (n-1)! \quad \text{for } n \in \mathbb{N}$$

We define p.d.f of x as,

$$f_X(x;\alpha,\beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} \quad x > 0$$

Notice that Gamma$(1,\beta) = \text{Exp}(\beta)$. That is, exponentional distribution is a specific case of gamma distribution. Gamma distribution itself is very useful and heavily used in this field. Moreover, we can derive more advanced (which I have difficulties understanding) distributions from the gamma function. For now, we will end our discussion here.

2.4 Multivariate Distribution

In practical word, we often work with multiple r.v in the same experiment. 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.
First, we define a special vector,

Definition 2.4.1. Let $X_1, X_2, .., X_n$ be r.vs. We call $X = \{X_1, X_2, ..., X_n\}$ **a random vector**.

Definition 2.4.2. If r.vs $X_1, X_2, ..., X_n$ are **independent** an have the same **marginal distribution** with c.d.f F, we define these r.vs as **independent and identically distributed**, shortly i. i. d, with notation,

$$X_1, ..., X_n \sim F$$

similarly, we show the p.d.f the same way. i.i.d property is very important in statistical field.

We can apply multivariate c.d.f as

Definition 2.4.3. For n r.v $\{X_1, X_2, .., X_n\}$, the multivariate c.d.f $F_{X_1,X_2,...,X_n}$ 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 c.d.f in general sense for n r.vs.

Similarly, we can define multivariate p.m.f as,

Definition 2.4.4. For random vector X,, the multivariate p.m.f $f_{X_1,X_2,...,X_n}$ is given by,

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

This is generalized form of p.m.f

Similarly, we define,

Definition 2.4.5. We know that c.d.f and p.d.f are related by derivative. Then, For random vector X, the multivariate p.d.f $f_{X_1,..,x_N}$ is given by,

$$f_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n) = \frac{\partial^n F_{X_1,..,X_n}(x_1, ..., x_n)}{\partial x_1 \partial x_2 ... \partial x_n}$$

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 $X_1, X_2, .., X_n$

Formally, **Marginal distribution is the probability of a single event (or r.v) occuring, independent of other events**. Therefore implementing marginal distributions are rather easy. In multivariate distributions, we redefine the needed variable as a "constant" and work with other variables only.

Definition 2.5.1. If X is a random vector with p.m.f $f_{X_1,X_2,..,X_n}$, then we define marginal distribution as,

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

Similarly,

Definition 2.5.2. We define marginal p.d.f as ,

$$f_{X_i}(x_i) = \int\int\int ... \int f(x_1, x_2, .., x_n)dx_1..dx_{i-1}dx_{i+1}...dx_n$$

Similarly, c.d.f follows the same rule. $F_X(x) = F(x, a, b, c, ...)$.
Remark: Marginality and conditionality are not the same thing. They look similiar, but their definitions are subtly different.

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,

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

The definition persists for multivariate distributions.

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 p.m.f f_{X_y}. THen X and Y are independent only and only if ,

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

The definiton persists for multivariate distributions.

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 p.m.f** as,

Definition 2.7.2. p.m.f of X conditional respect to Y can be written as

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

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/p.m.f f_X and c.d.f 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, p.m.f 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 c.d.f 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).

2.9 Exercises

1. Let $X \sim N(0, 1)$ and $Y = e^X$.
   1. plot p.d.f of X and Y.
   2. (The practical experiment) let x be a vector with 10000 numbers randomly selected from $N(0, 1)$, i .e $x = (x_1, .., x_{10000})$. Let y be a similiar vector where $y_i = e^{x_i}$, and plot the result of the histogram. Compare with the theorical $p.d.f$

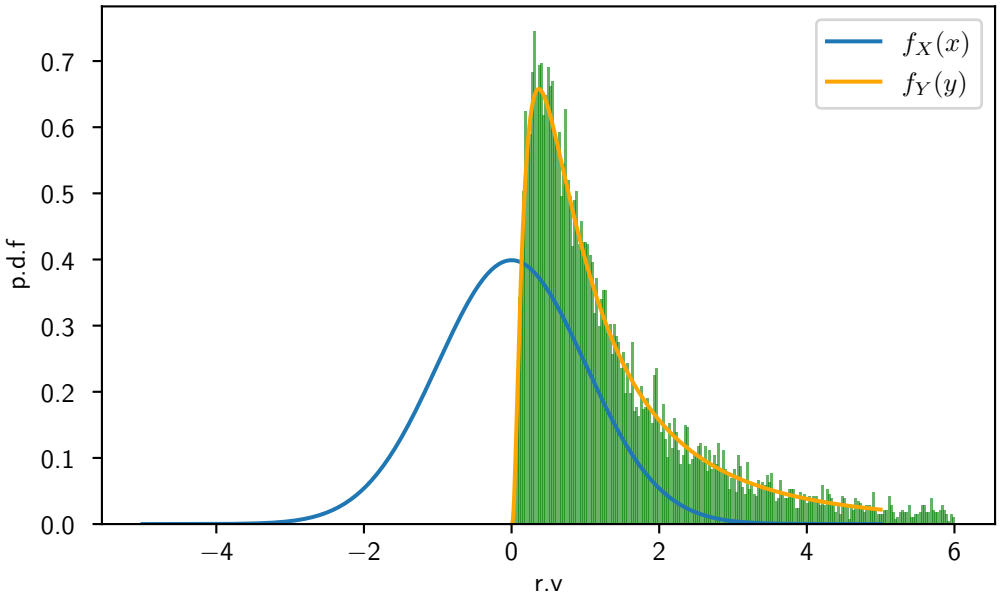
Solution1 Source
First, let's derive $f_Y(y)$.

$$\begin{aligned} F_Y(y) = P(Y \leq y) &= P(e^X \leq y) = P(X \leq \ln y) \\ &= F_X(\ln y) = \Phi(\ln y) \implies \\ f_Y(y) = \frac{\mathrm{d}\Phi(\ln y)}{\mathrm{d}y} &= \frac{\mathrm{d}\Phi(\ln y)}{\mathrm{d}\ln y}\frac{\mathrm{d}\ln y}{\mathrm{d}y} = \phi(\ln y)/y \end{aligned}$$

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import norm


# Generate an array with 1000 float elements, evenly spaced.
X = np.linspace(-5,5, num = 1000)
plt.figure(figsize=(6, 3.5))
# norm.pdf(x) gives the value of p.d.f of N(0,1) at x.
plt.plot(X, norm.pdf(X), label = "$f_X(x)$")
# We implement the equation we derived.
plt.plot(X, norm.pdf(np.log(X))/X, color = 'orange', label = "$f_Y(y)$" )
plt.title(r"Graph 1.")
plt.legend(loc='upper right')
plt.ylabel("p.d.f")
plt.xlabel("r.v")

# Now we draw the histogram plot
rng = np.random.default_rng() # initialize
totalPoints = 10000 # total number we will randomly select
arrayX = rng.normal(loc = 0, scale = 1, size = totalPoints) # array of selected
numbers
```

```
barSize = 200 # number of the bars
plt.hist(np.exp(arrayX), bins = barSize, range = (0,6),density = True, color =
'g', alpha = 0.6 )
plt.savefig('src/chapter2/fig1.pgf')
```

Graph 1.



It is kind of wrong to name y and x labels and ”p.d.f” and ”r.v”, but I really can’t think of better labels.
As expected, the green histogram nicely overlaps with our yellow plot.

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,

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

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,

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

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 $\mathrm{E}(X)$,

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

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,

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

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,

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

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

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

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

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 $\mathrm{E}(X) \geq 0$.

Proof. By definition of expectation, we have

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

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

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

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

$$\begin{aligned}\mathrm{E}\Big[a_1X_1 + a_2Y\Big] &= \int (a_1x + a_2y)dF_{X,Y}(x,y) \\ &= \int (a_1x)dF_X(x) + \int (a_2y)dF_Y(y) \\ &= a_1\int xdF_X(x) + a_2\int ydF_X(y) \\ &= a_1\mathrm{E}(X) + a_2\mathrm{E}(Y)\end{aligned}$$

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

$$\mathrm{E}\Big(\prod_{i=1}^{n} X_i\Big) = \prod_{i=1}^{n} \mathrm{E}(X_i)$$

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

$$\mathrm{E}(XY) = \int (xy)dF_{X,Y}(x,y) = \int xydF_X(x)dF_Y(y) = \int ydF_Y(y)\int xdF_X(x) = \mathrm{E}(X)\mathrm{E}(Y)$$

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,

$$\mathrm{E}(X|Y = y) = \int xdF_{X|Y}(x|y)$$

Note that $\mathrm{E}(X|Y)$ 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,

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

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

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

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 $\mu = \mathrm{E}(X)$. The Variance of X, denoted as $\mathrm{Var}(X)$ or $\mathrm{Var}(X)$ or σ^2 is the 2^{nd} central moment and is defined by,

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

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

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,

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

□

3.4 Conditional Variance

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

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

The conditional variance tells us how much of spread is left after We use $Y = y$. Reminder that $\mathrm{Var}(X|Y)$ 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,

$$\mathrm{Var}(Y) = \mathrm{E}\big[\mathrm{Var}(Y|X)\big] + \mathrm{Var}\big(\mathrm{E}[Y|X]\big)$$

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,

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

and **corelation** as,

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

Notice that $\mathrm{Cov}(X,X) = \mathrm{V}(X)$ 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

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

Proof. Similar to Variance one, we have,

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

□

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

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

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

Chapter 4

Statistical Inequalities

Statistical Inequalities provide a means of bounding measures and quantities, which is very useful practically and theorically. These inequalities are used for computations on Machine learning and AI, developing new methods and techniques for practical purposes and so on. IMO, International Mathematical Olympiads, also have a "Inequalities" section with questions that heavily use some of the methods and techniques soon be discussed here.

4.1 Probability inequalities

Probability inequaltiies are useful for boundign quantities that is hard to compute. Moreover, they are heavily used in the **theory of convergence**. Our first theory is,

Theorem 4.1.1. Let X be non-negative r.v and assume $\mathrm{E}(X)$ exists. Then **Markov's inequality** states that for any $a > 0$,

$$P(X \geq a) \leq \frac{\mathrm{E}(X)}{a}$$

Proof.

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} xf(x)dx = \int_{0}^{\infty} xf(x)dx \\ &= \int_{0}^{a} xf(x)dx + \int_{a}^{\infty} xf(x)dx \\ &\geq \int_{a}^{\infty} xf(x)dx \geq a\int_{a}^{\infty} f(x)dx \\ &= aP(X \geq a) \end{aligned}$$

□

Theorem 4.1.2. Let $\mu = E(X)$ and $\sigma^2 = V(X)$. **Chebyshev's inequality** states that,

$$P(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2} \quad \text{and} \quad P(|Z| \geq k) \leq \frac{1}{k^2}$$

Where Z is **standard score** i.e $Z = \frac{x-\mu}{\sigma}$.

Proof. The theorem is the direct consequence of the markov ineqaulity

$$P(|X - \mu| \geq a) = P(|X - \mu|^2 \geq a^2) \leq \frac{\mathrm{E}([X - \mu]^2)}{a^2} = \frac{\sigma^2}{a^2}$$

The second one is just a substituion. □

4.2 Expectation ineqaulities

Probably one of the most known ineqaulities of all time, that is also used in different field of mathematics, **Cauchy-Scwartz inequality** Simply states that,

Theorem 4.2.1. Cauchy-Scwarz ineqaulity states that,

$$\mathrm{E}^2(|XY|) \leq \mathrm{E}(X^2)\mathrm{E}(Y^2)$$

Another form of this theorem is,

$$\mathrm{Cov}^2(X,Y) \leq \sigma_X^2 \sigma_Y^2$$

This theorem is also known with its vector form. Consequently, this ineqaulity is very useful and practical, hence there are many unique and different proofs.
This ineqaulity is also popular on mathematical olympiads, with its familiar algebraic form.

Recall that a function f is **convex** if f is twice differentiable $f''(x) \geq 0$. Moreover f is **concave** if $-f$ is convex. Another very well known and popular ineqaulity,

Theorem 4.2.2. Jensen's ineqaultiy states that if g is convex,

$$\mathbb{E}g(X) \geq g(\mathbb{E}X)$$

If g is concave, the inequality symbol flips. Note that we used $\mathbb{E}X$ instead of $\mathrm{E}(X)$ for asthetical purposes.

Chapter 5

Convergence, CLT and LLN

5.1 Introduction

We already know the caclulus definition of convergence. We say that x_n **converges** to x if, for every $\epsilon > 0$.

$$|x_n - x| < \epsilon$$

as n goes to infinity.
However, there are multiple definitions of convergence in Probability and Statistics. The core idea is simple, in layman terms, as we repeat a process for a long time, something approaches to a thing. Mathematics requires regirousity, and hence there are multiple convergence definitions for the use case. We will learn convergence in the next section

CLT, in other words, **Central Limit theorem**, states that, under correct conditions, the distribution of **normalized verison** of the sample mean **converges in distribution** to a standard normal distribution. There are multiple types of CLT.

LLN, or **Law of Large Numbers** states that under right conditions, the sample average $\overline{X}$ **convergec in probability** to the expectation $\mu = \mathrm{E}(X)$. There are also types of LLN.

5.2 Types of Convergence

There are two main, most used types of convergence in the Statistics and Probability. The weakest one is, the **convergence in distribution**

Definition 5.2.1. Let $X_1, X_2, ...$ be a sequence of r.v with c.d.f $F_1, \ldots$. F_n is sait to be **converging in distribution** or **converge weakly** to c.d.f F of a r.v X if

$$\lim_{n\to\infty} F_n(x) = F(x)$$

For every x that F is **continous**. Convergence in distribution may be denoted as

$$X_n \xrightarrow{d} X$$

More stronger covnergence, **convergence in probability** is defined as,

Definition 5.2.2. Let $X_1, X_2, \ldots$ be a sequence of r.v. X_n is said to be **converging in probability** to r.v X if for all $\epsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0$$

Convergence in probabiltiy is denoted as,

$$X_n \xrightarrow{p} X$$

And other two less commonly used convergences,

Definition 5.2.3. Sequence $\{X_n\}$ converges **Almost surely** or **strongly** towards X if,

$$P\left(\lim_{n\to\infty} X_n = X\right) = 1$$

Almost surely convergence is denoted as,

$$X_n \overset{a.s}{\to} X$$

There is also stronger type of this version, called **sure convergence**. However, it is rarely used and there is no practical difference between this and the weaker version almost sure convergence. Therefore we don't talk about it.

and lastly, **convergence in mean**,

Definition 5.2.4. For a real number $r \geq 1$, X_n said to be **converges in r-th mean** to a r.v X if,

$$\lim_{x\to\infty} \mathrm{E}(|X_n - X|^r) = 0$$

This convergence is also called L_r convergence or L^r convergence and is denoted as,

$$X_n \overset{L^r}{\to} X$$

5.3 Properties of Convergences

Here is a basic diagram showing the chain of implications, from Wikipedia.

$$\begin{array}{ccccc} \xrightarrow{L^s} & & \underset{s>r\geq 1}{\Rightarrow} & & \xrightarrow{L^r} \\ & & & & \Downarrow \\ \xrightarrow{a.s} & \Rightarrow & \xrightarrow{p} & \Rightarrow & \xrightarrow{d} \end{array}$$

- Almost surely convergence implies convergence in probability

$$X_n \xrightarrow{a.s} X \quad \Rightarrow \quad X_n \xrightarrow{p} X$$

- Convergence in probability implies convergence in distribution

$$X_n \xrightarrow{p} X \quad \Rightarrow \quad X_n \xrightarrow{a.s} X$$

- Convergence in L^s implies convergence in L^r such that $s > r \geq 1$.

$$X_n \xrightarrow{L^s} X \quad \Rightarrow \quad X_n \xrightarrow{L^r} X$$

- Convergence in L^r implies convergence in probability.

$$X_n \xrightarrow{L^r} X \quad \Rightarrow \quad X_n \xrightarrow{p} X$$

Part II

Statistical Inference

Chapter 6

Estimation

Through the last chapter, we have talked about multiple probability distributions and their functions such as c.d.f, p.d.f, and p.m.f. However, we have assumed that we already knew the distribution and its properties. In practical world, it is not the case. We may try to find the average salary of the country, the fatality rate of a virus, and so on. Statistical inference, in shortly, is study of using the information sample we are given to deduce the characteristics of a population. Since majority of population is defined with **paramateres**, our investigation is mainly on finding, or estimating such paramteres.

6.1 Introduction

Majority of times we work with multiple parameters. However, we might be interested on only one of them. Therefore, we call the parameter of our interest to be **target parameter**.

Suppose we want to estimate μ of some experiment. We could give our estimate in two forms: **point estimate and interval estimate**. As name implies, that point estimate is a single value, however interval estimate is an interval.

Example 6.1.1. Suppose we want to estimate the average score μ the students will get from SAT score this year. We could use just a number 1240, a **point estimate**, or an interval $(1200, 1260)$, a **interval estimate**.

Examples for such estimators can be

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

Which esimates the mean μ. Notice that $\mathrm{E}(\overline{Y}) = \mu$ and,

$$S^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n}(X_i - \overline{X})^2\right)$$

Which estimates the variance σ^2. Notice that $\mathrm{E}(S^2) = \sigma^2$.
There can be multiple estimators estimating the same parameter. It is logical, since not all estimators are ideal for practical purposes because of quality of estimator.

6.2 Point Estimation

By convention we write the estimate of θ as $\widehat{\theta}$. Since θ is constant and by definition $\widehat{\theta}$ is a function, $\widehat{\theta}$ is a r.v. Remark that funtion of r.vs is a r.v. In more mathematical way,

Definition 6.2.1. Let $X_1, ..X_n \sim F$ be i.i.d. A point estimator $\widehat{\theta}$ is defined as,

$$\widehat{\theta} = g(X_1, \ldots, X_n)$$

We also define a very useful variable **bias** as,

$$\text{bias}(\widehat{\theta}, \theta) = \text{E}(\widehat{\theta}) - \theta$$

Here, θ is our target parameter, $\widehat{\theta}$ is the function we use to estimate our target, or the estimator. We usually write $\text{bias}(\widehat{\theta}, \theta) = \text{bias}(\widehat{\theta})$
Bias, in a literal sense, tells us the bias of the estimator we use. That is, the error that we may find when we estimate our parameter. We say that $\widehat{\theta}$ is **unbiased** if,

$$\text{bias}(\widehat{\theta}) = 0 \Rightarrow \text{E}(\widehat{\theta}) = \theta$$

We know that $\widehat{\theta}$ is a r.v. We call this r.v's distribution as **sampling distribution**. We also define, **standart error of** $\widehat{\theta}$ or standard deviation,

$$\sigma_{\widehat{\theta}} = \sqrt{\text{Var}(\widehat{\theta})}$$

It is logical to think that the estimator should converge (with more samples) to its target value, we define such property as,

Definition 6.2.2. If a point estimator $\widehat{\theta}$ converges to θ, we call that $\widehat{\theta}$ is **consistent**

With bias alone, we can't characterize the quality of the estimator. Because the values of $\widehat{\theta}$ may be far away than real value θ, but still be $\text{E}(\widehat{\theta}) = \theta$. Therefore, we also have to measure the variance in some way.
For such thing, we already have a tool,

Definition 6.2.3. The mean square error is defined as,

$$\text{MSE}(\widehat{\theta}) = \text{E}([\widehat{\theta} - \theta]^2)$$

in similiar fashion to the Variance definition, we can rewrite this equation as,

$$\text{MSE}(\widehat{\theta}) = \text{bias}^2(\widehat{\theta}) + \text{Var}(\widehat{\theta})$$

MSE is function of both its variance and bias, hence it is a better way of showing the quality of the estimator.

Example 6.2.1. Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. Let $\widehat{p} = n^{-1} \sum_{i=1}^{n} X_i$. We already know that $\text{E}(X) = p$ and $\text{Var}(X) = p(1-p)$. Our estimator is unbiased since,

$$\text{E}(\widehat{p}) = \frac{1}{n} \sum \text{E}(X_i) = p$$

Moreover, the estimator's variance is,

$$\sigma^2_{\widehat{p}} = \text{E}(\widehat{p}^2) - \text{E}(\widehat{p})^2 = p - p^2$$

We can intutively guess some estimators that could be effective for our purposes. But for many complex problems, it is not the case. We will learn new methods to calculate estimators in later chapters.

6.3 Confidence Intervals

Let's assume we are a scientist that want to measure the mean of the specific kind of mice's weight. It is unrealistic to measure **all of the mice**, hence we catch a small amount of them, probably in hundreds, measure them and gather the data in a datasheet. We **bootstrap** (we will learn what that term is in later chapters) the sample data, and find the sample mean. Now, we repeat the bootstrapping process thousands of times, which now we have a sample mean data.
Now, let's find numbers a, b such that 95% of our sample mean data resides in interval $[a, b]$. That is what **confidence interval** basically is.

Definition 6.3.1. Let X be a random vector. The $1-\alpha$ **confidence interval** for a parameter θ is an interval $[a, b]$ and $a = a(X)$, $b = b(X)$ functions such that,

$$P(a \leq \theta \leq b) \geq 1 - \alpha$$

Note that θ is unknown constant value, while a and b are random variables.
Taking the above example, $\alpha = 0.05$, which is a mathematical standard number used majority of time. $1 - \alpha$ is called **confidence coefficient**. We also call **lower and upper confidence limits** to a and b, sometimes also donated as θ_L and θ_U.

It is also possible to form *one sided confidence interval*, i.e,

$$P(\theta_L \leq \theta) \geq 1 - \alpha \qquad \text{or} \qquad P(\theta \leq \theta_U) \geq 1 - \alpha$$

The confidence intervals may be **closed or open**. For our purpose they are indifferent.

Chapter 7

Methods of Estimation (Parametric Inference)

In later chapter we have shortly talked about the point estimation. The estimator $\widehat{\theta}$ of a target parameter θ is a function of random variables of a sample, and therefore it itself is a random variable. The estimator has its own probability distribution, *sampling distribution*. We already know about *unbaised estimators* i.e $\mathrm{E}(\widehat{\theta}) = \theta$ and the *consistent estimator*. In this chapter, we will learn more deeply about the mathematical properties of the point estimators. Additionally, we will learn new methods to derive estimators, since until now we listened our intuition.

7.1 Properties of Point Estimation: Efficienty, Consistency, Sufficiency

Relative Efficiency

We already know that it is possible to have multiple estimators for one target parameter. We even learnt a new definiion, MSE to convey the quality of such estimators. If we have two unique and unbiased estimators $\widehat{\theta}_1$ and $\widehat{\theta}_2$, it is logical to pick the estimator that has the lowest variance, since the lower the MSE, more efficient the estimator is. To convey such idea, we use,

Definition 7.1.1. Given two unbaised estimators $\widehat{\theta}_1$ and $\widehat{\theta}_2$, then **the efficiency of $\widehat{\theta}_1$ relative to $\widehat{\theta}_2$**, denoted as $\mathrm{eff}(\widehat{\theta}_1, \widehat{\theta}_2)$, is defined as,

$$\mathrm{eff}(\widehat{\theta}_1, \widehat{\theta}_2) = \frac{\mathrm{Var}(\widehat{\theta}_2)}{\mathrm{Var}(\widehat{\theta}_1)}$$

Note that if $\mathrm{t}_1 \colon \mathrm{w}, \mathrm{t}_2$ is bigger than one, then it is true that $\widehat{\theta}_1$ is relatively more efficient than $\widehat{\theta}_2$.

Consistency

We have already talked about consistency before, we say the estimator is consistent of it converged to the target parameter,

Definition 7.1.2. The estimator $\widehat{\theta}$ of θ is consistent if for any positive number ϵ,

$$\lim_{n \to \infty} P(|\widehat{\theta}_n - \theta| \leq \epsilon) = 1$$

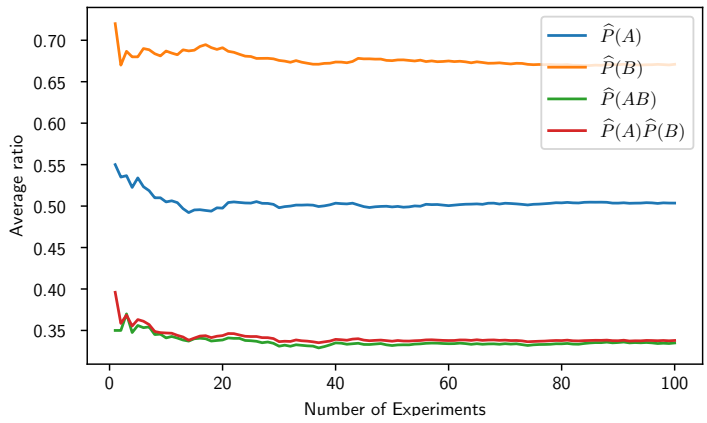
The graph (below) from latter exercises is also consistent, since visually it becomes a straight line where it equals to the target parameter.

Since consistent estimators converge to the target parameter, it is logical to think the variance also converges to 0. Think in a way that the graph of the estimator has to become straight from long wavy and curvy lines i.e it is direct consequence of convergence (real analysis stuff?). Indeed,

Theorem 7.1.1. The unbiased estimator $\widehat{\theta}$ of θ is a consistent estimator if,

$$\lim_{n\to\infty} \text{Var}(\widehat{\theta}_n) = 0$$

REMARK: There are multiple concepts of "convergence". In this case, it is **convergence in probability**. I really should study real analysis huh.



Sufficiency

We know that the value $\overline{X}$ (average value) is a unbiased estimator for mean μ of X. At this point, we no longer need the sample data to estimate the μ, since we can summarize the information just with the estimator $\overline{X}$. But, do the $\overline{X}$ retain all the information about X?. If it does, we call such estimator **sufficient**. That is all the sufficiency is for.

We can mathematically convery this property as conditional distribution of our sample data, given the estimator. If the distribution is dependent on our target parameter, it can't be sufficient. In more mathematical way,

Definition 7.1.3. A **statistic** is a function of data (Remark: all estimators are statistic but not all statistic are estimators). A statistic $U = t(X_1, .., X_n)$ of θ is sufficient if conditional distribution of $X_1, ..., X_n$ given U is not dependent on θ.

If conditional distribution is dependent on the target parameter, it is intuitive to think the statistic does not contain all the information.

Sufficiency is useful since it helps us to *assessing information on the entire population without the need of all the data*.
Say you get your grade on an exam and you want to know how well you did compared to your classmates. If you are given a sample mean and variance, you can do this without asking everyone's grades.

7.2 Method of Moments

Until now, we have used our intuiton to find estimators. For example, it is logical to think that $\overline{X}$ would be an ideal estimator for μ of X. However, in practical world we have to generate the parametric estimators more "mathematically". First, we introduce with a new simple definition,

Definition 7.2.1. k-th sample moment m_k is average of μ_k i.e

$$m_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

In section 3.1 we talked about **raw moments**. Raw moments convey the properties of the distribution i.e raw moments are some functions of the desired parameters. The first raw moment is the mean $\mu_1 = \mu$, the second raw moment is expression of variance $\mu_2 = \sigma^2 + \mu^2$ and so on.

The idea method of moment is we can use m_k as good estimator of μ_k, and from μ_k we can derive expressions for our target parameter.G

7.3 Method of Maximum Likelihood

The method of moments are very simple and intuitive, but it is unefficient. We have a better and sophisticated method called **method of maximum likelihood**. There is a great video by Josh Starmer that explains the method very well.

Assume that we have a sample data, and we want to estimate parameters of the distribution that describes the sample data. The idea is that we find such estimator that maximaze the **likelihood** of getting our sample data relative to the parameter.

Definition 7.3.1. The **Likelihood function** is defined as,

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(X_i; \theta)$$

Also we define **log-likelihood function** as,

$$\ell_n(\theta) = \log \mathcal{L}_n(\theta)$$

At last, we define the **maximum likelihood estimator** MLE denoted by $\widehat{\theta}_n$ as the value of θ that maximizes $\mathcal{L}_n(\theta)$, or better $\ell_n(\theta)$, since working with logs are easier that multiplicative functions for maximizing.

We already know that θ is a unknown constant we want to estimate. The $\mathcal{L}_n(\theta)$ describes the likelihood of each sample data, respect to θ. Since it is intuitive to maximize the likelihood (because the sample data is already happened and should be maximized), it should also estimate our value θ.

Chapter 8

Hypothesis Testing and p-value

8.1 Null and Alternate Hypothesis

The hypothesis testing is very similar to the scientific method. Scientists across the different fields use scientific method for their academical purposes. The observe, formulate a theory, experiment and test the theory. There is a similar method called **hypothesis testing** for statistical inference. First, we will introduce some notations and definitions,

Definition 8.1.1. Null Hypothesis, denoted by H_0, is the hypothesis to be tested.
Alternate Hypothesis, denoted by H_1, is the hypothesis contradictory to the null hypothesis. We usually try to support, since this way we could use *proof by contradiction*. If our evidence (data) favors the alternative hypothesis, we reject the null hypothesis. Formally, we wish to test,

$$H_0 : \theta \in \Theta_0 \qquad or \qquad H_1 : \theta \in \Theta_1$$

Where Θ_0 and Θ_1 are disjoint sets of parameter space Θ.

Example 8.1.1. Let $X_1, ..., X_n \sim N(\mu, \sigma^2)$ with known variance σ^2 and uknown mean μ. We wish to test the hypothesis,

$$H_0 : \mu = \mu_0 \qquad or \qquad H_1 : \mu \neq \mu_0$$

In order to test the hypothesis, it is logical to calculate $\overline{X}$ and compare it with μ_0. It is reasonable to reject H_0 if $\overline{X}$ is far away than μ_0. But how much far away exaclty?
Rejection or Critical Region is a set denoted by R to describe "how far away" the result can be. If the result is in the set R, we reject the hypothesis.

Definition 8.1.2. The **Critical Region** is defined as,

$$R = \left\{ x : T(x) > c \right\}$$

Here, c is called **critical value**, and T is a **test statistic** to help testing our hypothesis. The main question in hypothesis testing is find appropriate T and c.

In above example, $\overline{X}$ is our T. We will learn about finding c now.

There are 4 different possibilities we can conclude from our test. See the table for the brief introduction.
If H_0 is true, and we reject it, we call this error **type I error**. The probability of type I error is denoted as α. α is also called **size/level of the test**. Similarly, if H_1 is true and we keep the H_0, (or reject H_1) we call this error **type II error**. The probability of type II error is denoted as β. Summarize this in table,

	Retain Null	Reject Null
H_0 true	$(1-\alpha)$	type I error (α)
H_1 true	type II error (β)	power of the test $(1-\beta)$

Definition 8.1.3. The **power function** of a test with critical region R is a function that gives the **power of the hypothesis test**, that is $(1 - \beta)$, defined as,

$$\pi(\theta) = P(X \in R)$$

Some of the books notate this function as $B(\theta)$, power(θ) and even $\beta(\theta)$. **The power of the hypothesis test**, in literal sense, measures how powerful our hypothesis test is i.e the probability of rejecting H_0 while H_1 is true. It is commonly denoted as $1 - \beta$. See the table for more clarity.