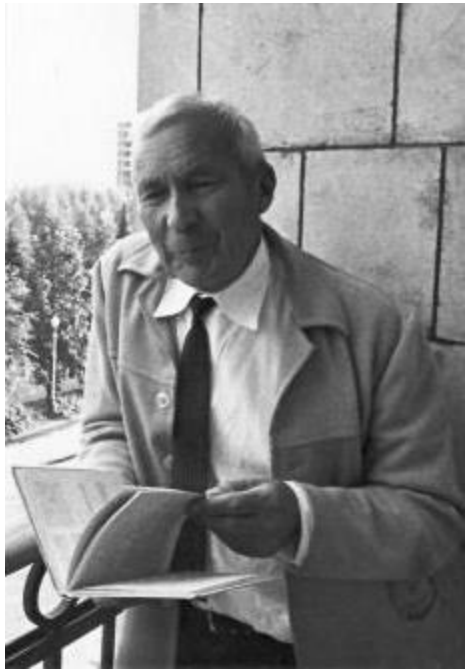
Math Primer
(15. Probability Theory)

Contents

- σ −algebra
- Filtration
- Probability Measure
- Random Variable
- Distribution
- Expectation
- Variance
- Borel σ −algebra



A. Kolmogorov
(1903-1907)

σ −algebra

- The study of probability begins with a 'random experiment'. For example, consider a random a coin toss experiment. Let's understand some very important definitions

Outcome of the experiment ω : $\omega = H$ or T

Sample space Ω (set of all possible outcomes) : $\Omega = \{H, T\}$

Event A (any sub-set of Ω) : e.g., $A = \phi$ or $A = \{H\}$

(We say Event A occurred if the outcome belongs to A , i.e., $\omega \in A$)

Similarly, if our experiment consists of two tosses, then

Outcome $\omega = HH$ or HT or TH or TT

Sample space $\Omega = \{HH, HT, TH, TT\}$

Event $A = \{HH, HT\}$ … first toss is head or $A = \{HH, HT, TH\}$ … at least one head

σ −algebra

- Consider a collection $\mathcal{F}$ of sub-sets of Ω (events) which satisfies the following properties

I. $\phi \in \mathcal{F}$

II. For every $A \subseteq \Omega$, $A \in \mathcal{F} \Longrightarrow A^C \in \mathcal{F}$

III. For every $A, B \subseteq \Omega$, $A, B \in \mathcal{F} \Longrightarrow A \cup B \in \mathcal{F}$

IV. If $\{A_n : 1,2,3 \ldots \}$ is a collection of sub-sets of Ω with $A_n \in \mathcal{F}$ for each $n >= 1$, then $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$

The last one basically says it's closed under countable union. If a collection of sub-sets of Ω satisfies the first three, then it's called an algebra in Ω. If it also satisfies the 4th condition, it's called a σ −algebra. The pair $(\Omega, \mathcal{F})$ is called a measurable space.

Exercise - 1. If $\Omega = \{1,2,3,4\}$ and $\mathcal{F} = \big\{\{1\}, \{2,3\}, \{4\}, \{2,3,4\}, \{1,2,3\}, \{1,4\}, \phi, \Omega\big\}$, is $\mathcal{F}$ a σ −algebra on Ω ?

Exercise - 2. Show that if A and B are in the σ −algebra, then $A \cap B$ and $A - B$ are also in the σ −algebra.

σ −algebra

- Consider a random experiment consisting of tossing a coin 3 times. The sample space is given as

$$\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$$

Some important σ −algebra are

$$\mathcal{F}_0 = \{\phi, \Omega\}$$

$$\mathcal{F}_1 = \{\phi, \Omega, \{HHH, HHT, HTH, HTT\}, \{THH, THT, TTH, TTT\}\}$$

$\mathcal{F}_2 = \{\phi, \Omega, \{HHH, HHT\}, \{HTH, HTT\}, \{THH, THT\}, \{TTH, TTT\},$

and all sets that can be built taking unions of these $\}$

$\mathcal{F} = \mathcal{F}_3 = \{$all subsets of $\Omega\}$ ... power set

σ −algebra

- Let's simplify the notations by defining some events

$$A_H = \{HHH, HHT, HTH, HTT\} \text{ ... head in the first toss}$$
$$A_T = \{THH, THT, TTH, TTT\} \text{... tail in the first toss}$$

$$\mathcal{F}_1 = \{\phi, \Omega, A_H, A_T\}$$

We interpret σ −algebra as record of information. Suppose an outcome(ω) has happened which we don't know, but we are told if each of the event in $\mathcal{F}_1$ is true or false i.e., whether ω belongs to each elements of $\mathcal{F}_1$ or not. This will lead to the precise knowledge of the first toss. In other words, $\mathcal{F}_1$ stores the information till the first toss.

For example, if the actual outcome(ω) is HHT which we do not know. But we will say $\omega \notin \phi$, $\omega \in \Omega$, $\omega \in A_H$, $\omega \notin A_T$. This information tells us that the first toss was H, nothing more.

σ −algebra

- Similarly, if we define

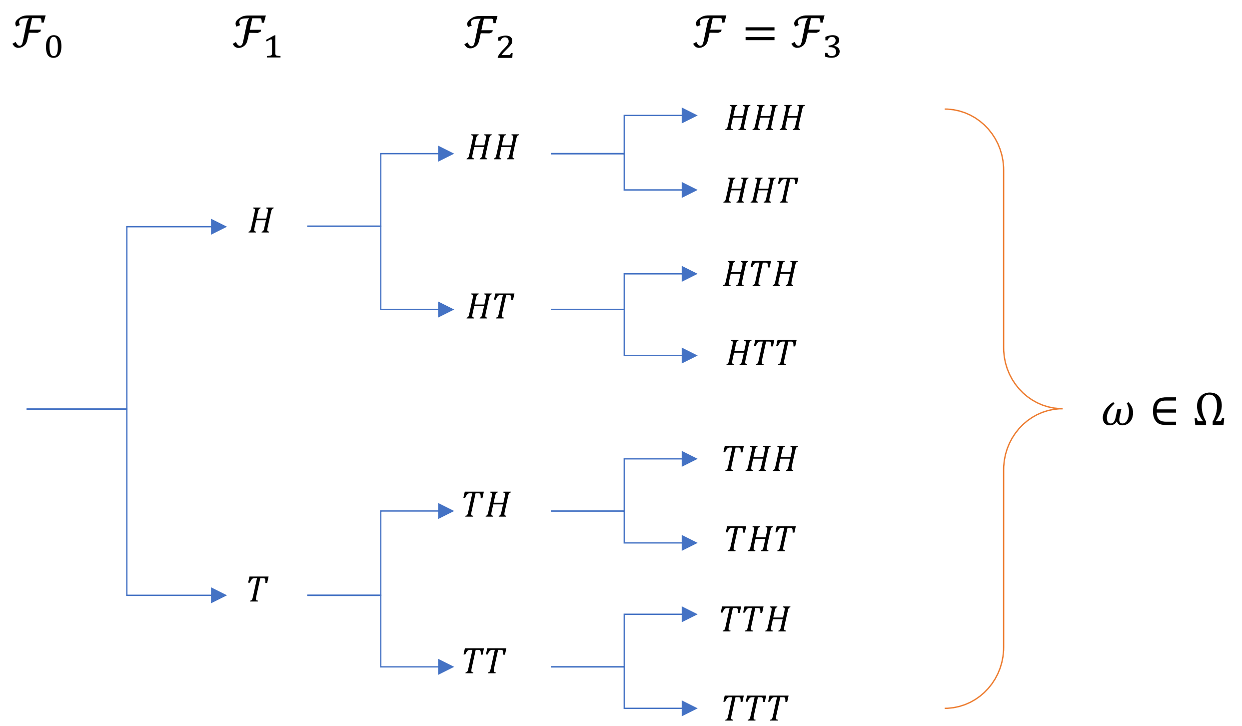
$$A_{HH} = \{HHH, HHT\} \text{ ... HH in the first two tosses}$$
$$A_{HT} = \{HTH, HTT\} \text{ ... HT in the first two tosses}$$
$$A_{TH} = \{THH, THT\} \text{ ... TH in the first two tosses}$$
$$A_{TT} = \{TTH, TTT\} \text{ ... TT in the first two tosses}$$

$$\mathcal{F}_2 = \{\phi, \Omega, A_{HH}, A_{HT}, A_{TH}, A_{TT},$$
$$A_H, A_T, A_{HH} \bigcup A_{TH}, A_{HH} \bigcup A_{TT}, A_{HT} \bigcup A_{TH}, A_{HT} \bigcup A_{TT},$$
$$A_{HH}{}^c, A_{HT}{}^c, A_{TH}{}^c, A_{TT}{}^c\}$$

We can see that $\mathcal{F}_2$ records the information till the first two coin tosses.

Filtration

- The σ −algebra $\mathcal{F} = \mathcal{F}_3$ which is the collection of all subsets of Ω (Power Set of Ω) contains full information.
- A filtration is a sequence of σ −algebras $\mathcal{F}_0, \mathcal{F}_1, \dots, \mathcal{F}_n$ such that $\mathcal{F}_i \supset \mathcal{F}_{i-1}$. So, basically all σ −algebra in the sequence contain all the elements of all previous σ −algebras.



Probability Measure

- We understand probability as the chances of an event happening. Formally speaking, a probability measure $\mathbb{P}$ defined on $(\Omega, \mathcal{F})$ is a function that maps $\mathcal{F}$ to $[0,1]$ and has the following properties

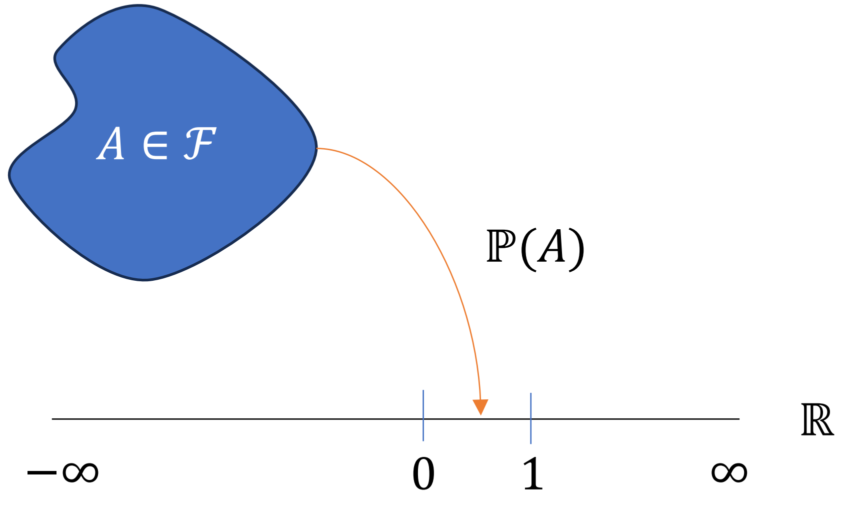
I. $\mathbb{P}(\phi) = 0$

II. $\mathbb{P}(A) \geq 0$ for every $A \in \mathcal{F}$

III. $\mathbb{P}(\Omega) = 1$

IV. if $A_1, A_2, \ldots,$ is a sequence of disjoint sets in $\mathcal{F}$, then

$$\mathbb{P}\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} \mathbb{P}(A_k)$$



The triplet $(\Omega, \mathcal{F}, \mathbb{P})$ is called a probability-space

Probability Measure

- Example – Assume the probability of H and T in one coin toss be 1/3 and 2/3 respectively, then we can derive the probabilities of all complex events.
- First, we find out the probabilities of all the singletons of Ω. Here we assume all tosses are independent

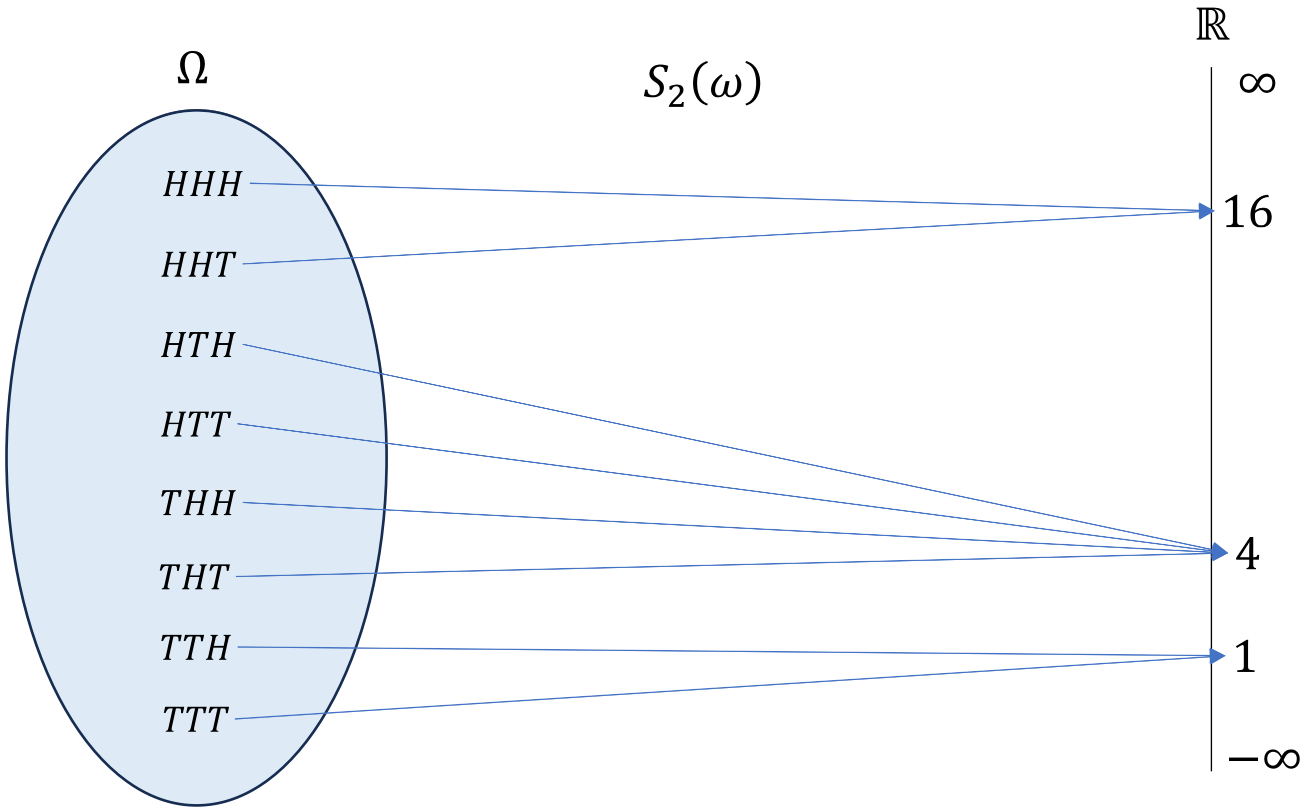
$$\mathbb{P}\{HHH\} = \left(\frac{1}{3}\right)^3, \mathbb{P}\{HHT\} = \left(\frac{1}{3}\right)^2\left(\frac{2}{3}\right), \mathbb{P}\{HTH\} = \left(\frac{1}{3}\right)^2\left(\frac{2}{3}\right), \mathbb{P}\{HTT\} = \left(\frac{1}{3}\right)\left(\frac{2}{3}\right)^2$$
$$\mathbb{P}\{THH\} = \left(\frac{2}{3}\right)\left(\frac{1}{3}\right)^2, \mathbb{P}\{THT\} = \left(\frac{2}{3}\right)^2\left(\frac{1}{3}\right), \mathbb{P}\{TTH\} = \left(\frac{2}{3}\right)^2\left(\frac{1}{3}\right), \mathbb{P}\{TTT\} = \left(\frac{2}{3}\right)^3$$

Note that the probability of H in the first toss in an experiment with 3 tosses can be recovered as follows. We use the property IV to add the probabilities of disjoint sets

$$\mathbb{P}(A_H) = \mathbb{P}\{HHH, HHT, HTH, HTT\} = \mathbb{P}(\{HHH\} \cup \{HHT\} \cup \{HTH\} \cup \{HTT\})$$
$$= \left(\frac{1}{3}\right)^3 + \left(\frac{1}{3}\right)^2\left(\frac{2}{3}\right) + \left(\frac{1}{3}\right)^2\left(\frac{2}{3}\right) + \left(\frac{1}{3}\right)\left(\frac{2}{3}\right)^2 = \frac{1}{3}$$

Random Variable

- Let's consider a wealth process (S). The initial wealth is $S_0 = \$4$. Now with every coin toss the wealth doubles if the outcome is H and halves if the outcome is T. In this process, S_1, S_2 and S_3 are random variables. A random variable is function that is a mapping from $\Omega \longrightarrow \mathbb{R}$
- For example, consider the random variable $S_2(\omega)$



What is the pre-image under S_2 of this interval [4,20] ?

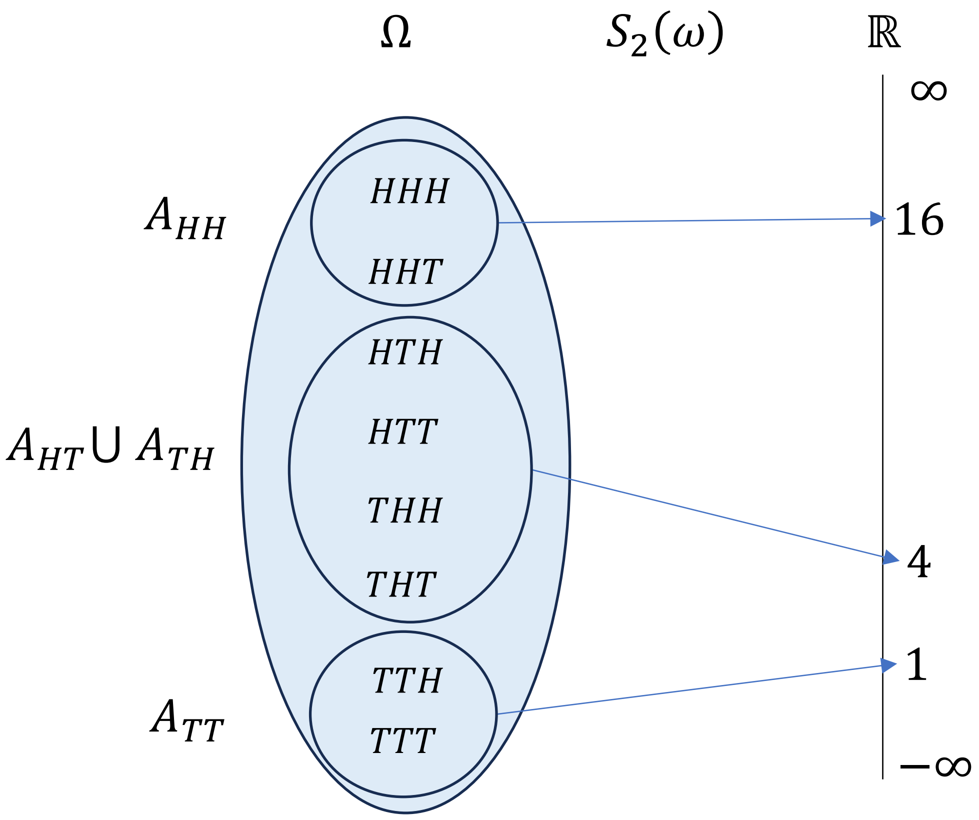
σ −algebra generated by a Random Variable

- The complete collection of pre-images under S_2 is given as $\phi, \Omega, A_{HH}, A_{TT}, A_{HT} \cup A_{TH}$ and all possible unions of these. This collection of sets is a σ −algebra and is called the σ −algebra generated by random variable S_2 and is denoted by $\sigma(S_2)$

$$\sigma(S_2) = \{\phi, \Omega, A_{HH}, A_{TT}, A_{HT} \bigcup A_{TH}, A_{HH} \bigcup A_{TT}, {A_{HH}}^c, {A_{TT}}^c\}$$

- The above σ −algebra contains the same information that we will get by observing S_2. For example, if we are told that the outcome is not in A_{HH} or A_{TT} but in $A_{HT} \cup A_{TH}$ then we would know that the first two tosses contain one H and one T, nothing more. The same information is contained in saying $S_2 = 4$.

- Note that $\sigma(S_2) \neq \mathcal{F}_2$ as $\mathcal{F}_2$ contains more information. In $\mathcal{F}_2$, we have precise information of the first two tosses.



Measurable

- Let $\mathcal{F}$ be the σ −algebra of all subsets of Ω. Let X be a random variable $\Omega \longrightarrow \mathbb{R}$ and $\sigma(X)$ be the σ −algebra generated by X. We say X is $\mathcal{G}$ − measurable if $\sigma(X) \subseteq \mathcal{G}$

  For Example, we can say S_2 is $\mathcal{F}_2$ measurable but not $\mathcal{F}_1$ measurable

This simply means that if we know $\mathcal{F}_2$, we precisely know the value S_2

But if we know only $\mathcal{F}_1$, then S_2 is still random

Induced Measure

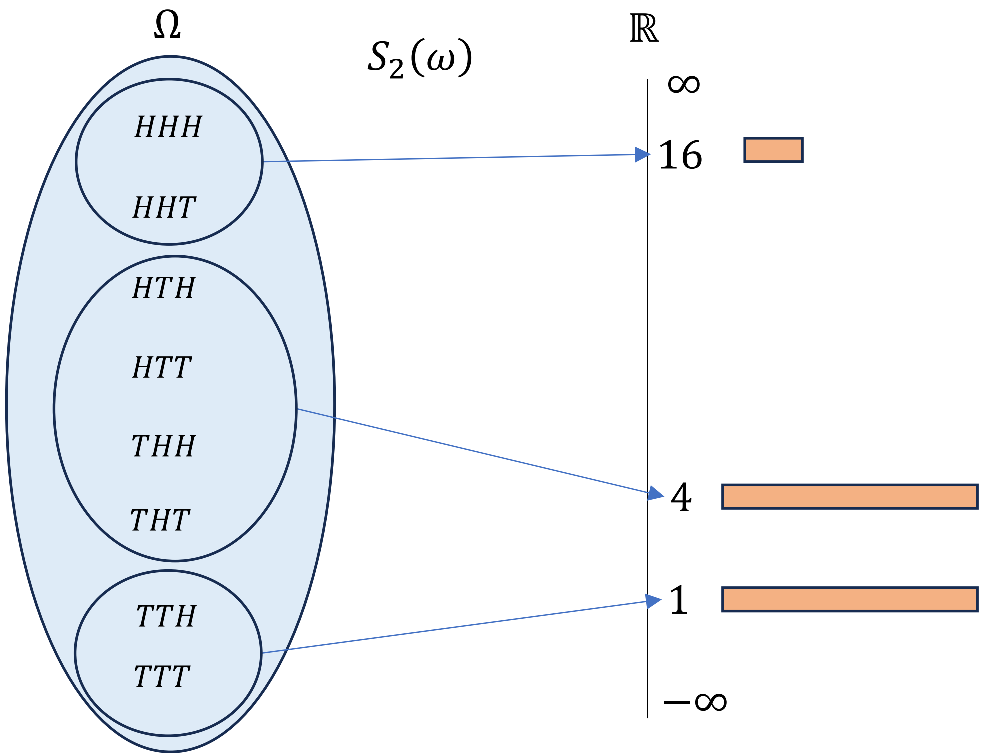
- Consider the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a random variable $X: \Omega \longrightarrow \mathbb{R}$. Given a set $A \subseteq \mathbb{R}$, we define the induced measure of A to be $\mathcal{L}_X(A) = \mathbb{P}(X \in A)$

In other words, it tells us the probability that X takes value in A. For example, consider our random variable S_2. We can place probability masses on $\mathbb{R}$ based on the induced measure.

$\mathbb{P}(A_{HH}) = \frac{1}{9}$

$\mathbb{P}(A_{HT} \cup A_{TH}) = \frac{4}{9}$

$\mathbb{P}(A_{TT}) = \frac{4}{9}$



$\mathcal{L}_{S_2}(\phi) = \mathbb{P}(\phi) = 0$

$\mathcal{L}_{S_2}(\mathbb{R}) = \mathbb{P}(\Omega) = 1$

$\mathcal{L}_{S_2}[0, \infty) = \mathbb{P}(\Omega) = 1$

$\mathcal{L}_{S_2}[0,3] = \mathbb{P}(S_2 = 1) = \mathbb{P}(A_{TT}) = \frac{4}{9}$

Probability Distribution

- There are two main ways of recording the information $\mathcal{L}_X$

Probability mass function $\coloneqq f_X(x)$

Cumulative distribution function $\coloneqq F_X(x)$

In our example,

$$f_{S_2}(x) = \begin{cases} \frac{4}{9}, & x = 1 \\ \frac{4}{9}, & x = 4 \\ \frac{1}{9}, & x = 16 \end{cases} \qquad F_{S_2}(x) = \begin{cases} 0, & x < 1 \\ \frac{4}{9}, & 1 \leq x < 4 \\ \frac{8}{9}, & 4 \leq x < 16 \\ 1, & x \geq 16 \end{cases}$$

Note that random variables and probabilities are separate constructs and should not be mixed.

Expected Value

- The expected value of a random variable X is given as

$$\mathbb{E}[X] = \sum_{\omega \in \Omega} X(\omega)\mathbb{P}(\omega)$$

The sum is over all outcomes in Ω. Since in our example, X only takes discrete values $(x_1, x_2, \ldots, x_k)$ we can rewrite expectation to be a sum over the real number line $\mathbb{R}$

$$\mathbb{E}[X] = \sum_{k=1}^{n} x_k \mathcal{L}_X\{x_k\}$$

In our example, $\mathbb{E}[S_2] = 1 * \mathcal{L}_{S_2}\{1\} + 4 * \mathcal{L}_{S_2}\{4\} + 16 * \mathcal{L}_{S_2}\{16\} = 1 * \frac{4}{9} + 4 * \frac{4}{9} + 16 * \frac{1}{9} = 4$

Variance

- The variance of a random variable X is given as

$$\mathbb{V}[X] = \sum_{\omega\in\Omega}(X(\omega) - \mathbb{E}[X])^2\mathbb{P}(\omega)$$

Once again sum is over all outcomes in Ω. We can rewrite expectation to be a sum over the real number line $\mathbb{R}$

$$\mathbb{V}[X] = \sum_{k=1}^{n}(x_k - \mathbb{E}[X])\mathcal{L}_X\{x_k\}$$

In our example, $\mathbb{V}[S_2] = (1-4)^2 * \mathcal{L}_{S_2}\{1\} + (4-4)^2 * \mathcal{L}_{S_2}\{4\} + (16-4)^2 * \mathcal{L}_{S_2}\{16\}$

$$= 9 * \frac{4}{9} + 0 * \frac{4}{9} + 144 * \frac{1}{9} = 4 + 0 + 16 = 20$$

Borel σ −algebra $\mathcal{B}(\mathbb{R})$

PEAKS2TAILS

- The Borel σ −algebra is the smallest σ −algebra containing all open intervals in $\mathbb{R}$. It is denoted as $\mathcal{B}(\mathbb{R})$. The sets in $\mathcal{B}(\mathbb{R})$ are called Borel sets.

- Every open interval (a, b) is in $\mathcal{B}(\mathbb{R})$
- Every union of open intervals is in $\mathcal{B}(\mathbb{R})$ … property of σ −algebra
- All half-open lines are in $\mathcal{B}(\mathbb{R})$, e.g., $(a, \infty) = \cup_{n=1}^{\infty}(a, a+n)$, $(-\infty, a) = \cup_{n=1}^{\infty}(a-n, a)$
- The union $(-\infty, a) \cup (b, \infty)$ is also in $\mathcal{B}(\mathbb{R})$
- This means that every closed interval is also in $\mathcal{B}(\mathbb{R})$, $[a, b] = \big((-\infty, a) \cup (b, \infty)\big)^c$

It can be shown that closed-half lines, half-closed intervals, half-open intervals are also in Borel. Every set containing countably finite or infinite numbers are also in Borel.

Pretty much any imaginable set in $\mathbb{R}$ is in $\mathcal{B}(\mathbb{R})$. However, not all sets are Borel.

Borel Measurable

- Consider a function $f : \mathbb{R} \longrightarrow \mathbb{R}$. We say f is Borel-measurable if whenever $B \in \mathcal{B}(\mathbb{R})$, $f^{-1}(B) \in \mathcal{B}(\mathbb{R})$. Indirectly, we are saying that the σ −algebra generated by f should be in $\mathcal{B}(\mathbb{R})$ for the function to be Borel-measurable.

In practice, it's hard to produce functions which are not Borel-measurable. We will assume all sub-sets of $\mathbb{R}$ are Borel sets and all $f : \mathbb{R} \longrightarrow \mathbb{R}$ are Borel-measurable functions.

Borel Measurable

- Example – Consider the indicator function $\mathbb{I}_A(x)$ where $A \subseteq \mathbb{R}$ given as

$$\mathbb{I}_A(x) = \begin{cases} 1, & x \in A \\ 0, & x \notin A \end{cases}$$

Now if we take any $B \in \mathcal{B}(\mathbb{R})$,

$$\mathbb{I}_A{}^{-1}(x) = \begin{cases} \phi, & 0 \notin B, 1 \notin B \\ A, & 0 \notin B, 1 \in B \\ A^c, & 0 \in B, 1 \notin B \\ \mathbb{R}, & 0 \in B, 1 \in B \end{cases}$$

We can see $\{\phi, A, A^c, \mathbb{R}\}$ is a σ –algebra contained in $\mathcal{B}(\mathbb{R})$, so $\mathbb{I}_A(x)$ is Borel-measurable.

Borel Measurable

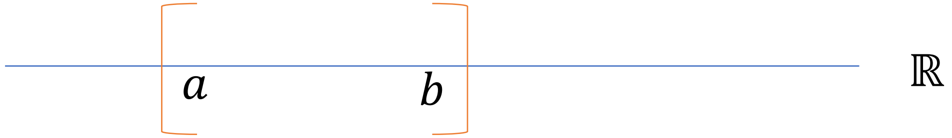
Exercise - 1. What is $\mathbb{E}[\mathbb{I}_A(x)]$ **?**

Exercise - 2. Examine the Borel-measurability for the indicator function $\mathbb{I}_{x \geq 100}$

Exercise – 3. Are there any functions which are measurable on $(\phi, \mathbb{R})$?

Measure Theory

- How to measure the subsets of real number line $\mathbb{R}$?



The first measure that comes to mind is the length $:= b - a$

In $\mathbb{R}^2$ and $\mathbb{R}^3$, this extends to the idea of Area and Volume

We can call this measure as a generalized volume (space captured by the interval)

- A Measure defined on $\left(\mathbb{R}, \mathcal{B}(\mathbb{R})\right)$ is a function $\mu: \mathcal{B} \longrightarrow [0, \infty]$ with the following properties

I. $\mu(\phi) = 0$

II. if $A_1, A_2, \dots,$ is a sequence of disjoint sets in $\mathcal{B}$, then $\mu\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} \mu(A_k)$

$(\Omega, \mathcal{F}, \mu)$ is called a measure-space

Measure Theory - Examples

PEAKS2TAILS

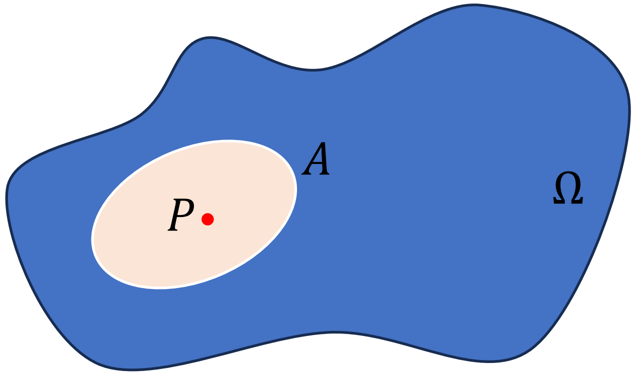
- Consider a measure space $(\Omega, \mathcal{F}, \mu)$. We measure $A \in \mathcal{F}$

1. Counting measure :

$$\mu(A) = \begin{cases} \neq A, & \text{if } A \text{ is a finite set} \\ \infty, & \text{if } A \text{ is an infinite set} \end{cases}$$

2. Dirac Measure

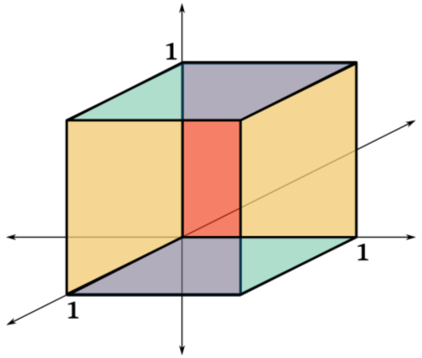
$$\delta_P(A) = \begin{cases} 1, & \text{if } P \in A \\ 0, & \text{if } P \notin A \end{cases}$$

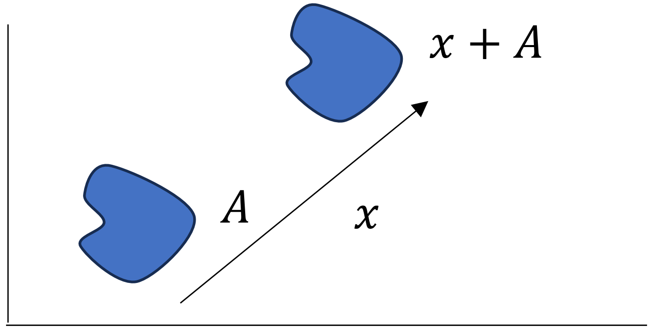


Measure Theory - Examples

- Counting Measure and Dirac measure are basic measures but are not generalized volume measures. We seek a normal volume measure in $\mathbb{R}^n$ which has the following 2 desirable properties

1. $\mu([0,1]^n) = 1$, Volume of a unit cube
2. $\mu(x + A) = \mu(A)$, invariant under translation





A measure has all the properties of Probability except that total measure of the space is not 1.

If a measure satisfies the above properties, it's called a Lebesgue measure

Lebesgue Measure

A Lebesgue measure is defined to be a measure on $\left(\mathbb{R}, \mathcal{B}(\mathbb{R})\right)$ which assigns the measure of each interval to be its length. Let's denote it as μ_0

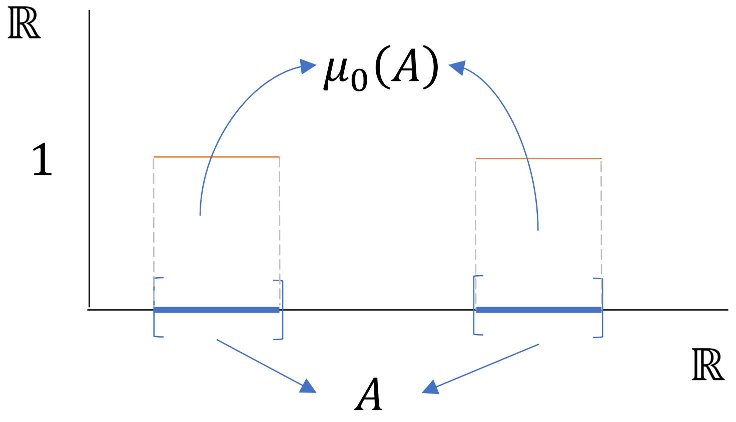
- Note that $\mu_0(\mathbb{R}) = \infty$
- The Lebesgue measure of a set containing only one point is 0 , i.e., $\mu_0\{a\} = 0$
- The Lebesgue measure of a countably infinite set is also 0 , i.e., $\mu_0\{a_1, a_2, \dots\} = 0$
- For uncountably infinite number of points, Lebesgue measure is can be 0, positive finite or infinite. We cannot simply add the measure for all uncountable points. Integration comes to the rescue !!

Lebesgue Integral – Indicator Functions

Consider the indicator function $\mathbb{I}_A(x)$ where $A \subseteq \mathbb{R}$ given as

$$\mathbb{I}_A(x) = \begin{cases} 1, & x \in A \\ 0, & x \notin A \end{cases}$$

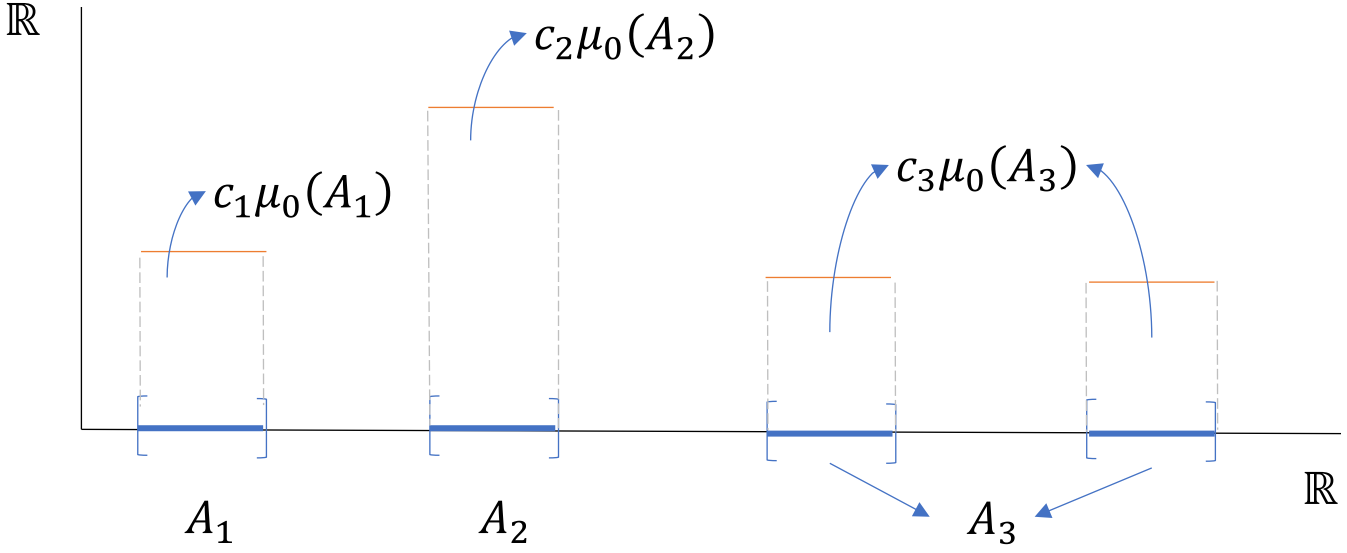
The Lebesgue integral of this function over $\mathbb{R}$ is $\int_{\mathbb{R}} \mathbb{I}_A(x) d\mu_0 = \mu_0(A)$



Lebesgue Integral – Simple Functions

A simple function takes a finite set of values and can be expressed as a linear combination of indicator functions (e.g., step function). In the following, h is a simple function

$$h(x) = \sum_{i=1}^{n} c_i \, \mathbb{I}_{A_i}(x) \quad \text{where } A_1, A_2, \dots, A_n \in \mathcal{B}(\mathbb{R}) \ , \ c_1, c_2, \dots, c_n \in \mathbb{R}$$



Lebesgue Integral of the simple function is

$$\int_{\mathbb{R}} h(x) d\mu_0 = \sum_{i=1}^{n} c_i \, \mu_0(A_i)$$

Lebesgue Integral – Non-Negative Simple functions

Let f be a non-negative simple function defined on $\mathbb{R}$, possibly taking value ∞ at certain points. If the Lebesgue integral $= \infty$, we say it's not integrable. If it's $< \infty$, we say it's integrable.

If f is non-negative, then $f(x) = \sum_{i=1}^{n} c_i \, \mathbb{I}_{A_i}(x)$, $c_1, c_2, \dots, c_n \geq 0$

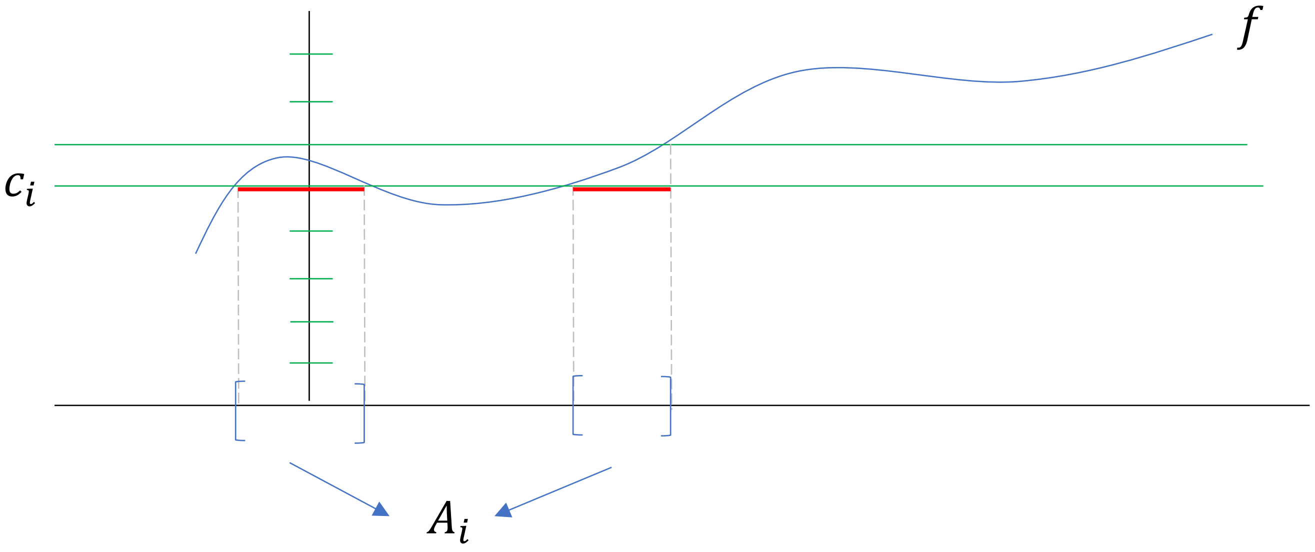
The Lebesgue integral of f is $\int_{\mathbb{R}} f(x) d\mu_0 = \sum_{i=1}^{n} c_i \, \mu_0(A_i) \in [0, \infty]$. It satisfies some basic properties

1. Linearity : $\int_{\mathbb{R}} (\alpha f + \beta g) d\mu_0 = \alpha \int_{\mathbb{R}} f d\mu_0 + \beta \int_{\mathbb{R}} g d\mu_0$
2. Monotonicity : $\int_{\mathbb{R}} f d\mu_0 \leq \int_{\mathbb{R}} g d\mu_0$ if $f \leq g$

Lebesgue Integral – Any function

We saw the Lebesgue integral of non-negative simple functions. What about functions that are not simple ? Let's generalize the Lebesgue integral of any function.

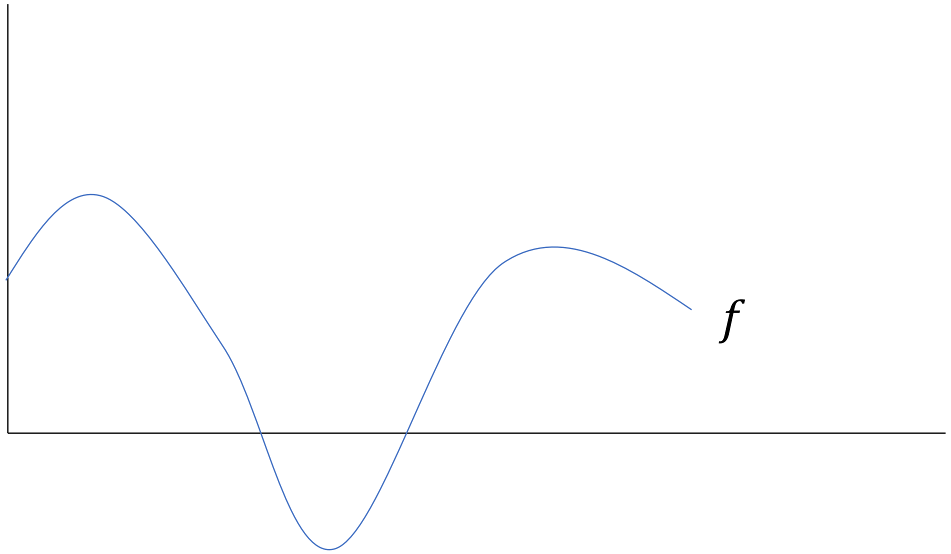
We can think of approximating the function f with a simple function h by considering many bands on Y-axis. For each interval A_i, we choose c_i, such that $f(x) \geq h(x)$, $x \in A_i$



We define Lebesgue integral of f as $\int_{\mathbb{R}} f d\mu_0 = \sup\left\{\int_{\mathbb{R}} h d\mu_0 \text{, where h is simple and } h(x) \leq f(x) \text{ for } x \in \mathbb{R}\right\}$

Lebesgue Integral – Any function

What if functions take negative values ? We can still only restrict ourselves to positive functions by partitioning the interval in positive and negative regions. Only take the absolute value while integrating and then subtract the negative part from the positive part.



$$f^+ = \max\{f(x), 0\}$$

$$f^- = \max\{-f(x), 0\}$$

$$\int_{\mathbb{R}} f(x) d\mu_0 = \int_{\mathbb{R}} f^+(x) d\mu_0 - \int_{\mathbb{R}} f^-(x) d\mu_0$$

Pointwise convergence

Consider $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu_0)$ measure space, and a sequence of functions $f_1 \leq f_2 \leq f_3 \leq \cdots \leq f_n$ converging to pointwise to a function limiting function $f(x)$ This is called pointwise convergence which basically means . $\lim_{n \to \infty} f_n(x) = f(x)$

e.g., consider a sequence of functions which are normal densities with mean n=1,2,... and variance 1. These functions converge to $f(x) = 0$

$$f_n(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-n)^2}{2}} \qquad \lim_{n \to \infty} f_n(x) \simeq f(x) = 0$$

Note $\int_{\mathbb{R}} f_n(x) d\mu_0 = 1$ for all n, so $\lim_{n \to \infty} \int_{\mathbb{R}} f_n(x) d\mu_0 = 1$, but $\int_{\mathbb{R}} f(x) d\mu_0 = 0$

(Integral of the limiting function is not equal to limit of the integral)

Exercise

PEAKS2TAILS

Find the limiting function of the following sequence and evaluate the limit of the integral of the function and the integral of the limiting function.

$$f_n(x) = \sqrt{\frac{n}{2\pi}}\, e^{-\frac{x^2}{2/n}}$$

Fatou's Lemma

Let f_n , $n = 1,2, \ldots$ be a sequence of nonnegative functions converging pointwise to f, then

$$\int_{\mathbb{R}} f d\mu_0 \leq \lim_{n \longrightarrow \infty} \int_{\mathbb{R}} f_n d\mu_0$$

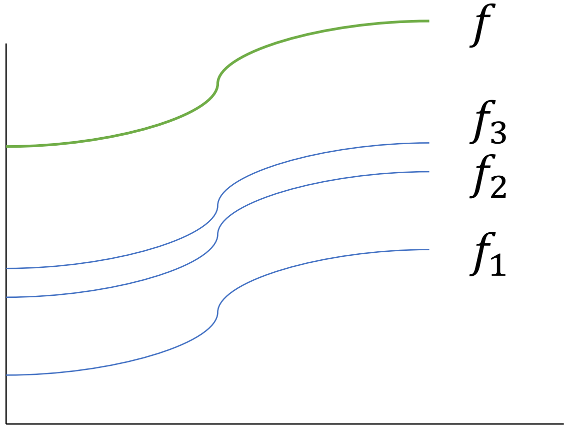
Note that in the previous two examples, we had $\int_{\mathbb{R}} f d\mu_0 = 0$ and $\lim_{n \longrightarrow \infty} \int_{\mathbb{R}} f_n d\mu_0 = 1$

Fatou's lemma only assumes functions are non-negative and the result is not too interesting either.

Monotone Convergence Theorem

Let f_n , $n = 1,2, \ldots$ be a sequence of nonnegative functions converging pointwise to f, and assume that $0 \leq f_1(x) \leq f_2(x) \leq \cdots$ for every $x \in \mathbb{R}$, then

$$\int_{\mathbb{R}} f \, d\mu_0 = \lim_{n \to \infty} \int_{\mathbb{R}} f_n \, d\mu_0$$



Both sides are allowed to be ∞

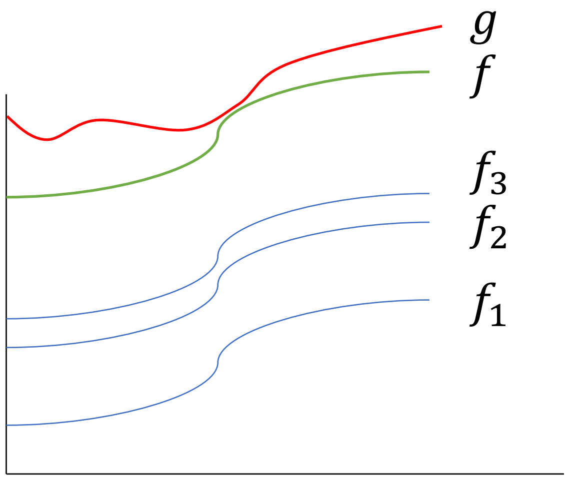
This is a powerful result as we can take the limit inside of the integral

Dominated Convergence Theorem

Let f_n , $n = 1,2,\dots$ be a sequence of functions which can be positive or negative converging pointwise to f, and assume there is a non-negative integrable function g(integrable majorant)

such that

$$|f(x)| \leq g(x) \quad \text{for all } x \in \mathbb{R}$$



Then

$$\int_{\mathbb{R}} f \, d\mu_0 = \lim_{n \longrightarrow \infty} \int_{\mathbb{R}} f_n \, d\mu_0$$

And both sides will be finite.

Probability as a Lebesgue Integral

Probability is a Lebesgue integral of the density function $\varphi(x)$

$$\mathbb{P}(A) = \int_A \varphi(x)\, d\mu_0$$

We can write $\varphi(x) = \dfrac{d\mathbb{P}}{d\mu_0}$ (Radon-Nikodym derivative of $\mathbb{P}$ w.r.t μ_0)

The density function $\varphi(x)$ is a non-negative function satisfying $\int_{\mathbb{R}} \varphi(x) d\mu_0 = 1$

For a standard uniform distribution, the probability measure = Lebesgue measure $\mathbb{P}(A) = \mu_0(A)$

Integrals in Probability Space

Integrals in Probability space is constructed using the same steps as a Lebesgue integral. Consider the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The integration of a random variable w.r.t the probability measure is called Expectation.

1. If X is an indicator, then $X(\omega) = \mathbb{I}_A(\omega) = \begin{cases} 1, & \omega \in A \\ 0, & \omega \notin A \end{cases}$, $\int_\Omega X \, d\mathbb{P} = \mathbb{P}(A)$
2. If X is a simple function, then $X(\omega) = \sum_{i=1}^{n} c_i \, \mathbb{I}_{A_i}(\omega)$, $\int_\Omega X \, d\mathbb{P} = \sum_{i=1}^{n} c_i \int_\Omega \mathbb{I}_{A_i} \, d\mathbb{P} = \sum_{i=1}^{n} c_i \, \mathbb{P}(A_i)$
3. If X is a non-negative general function

$$\int_\Omega X \, d\mathbb{P} = \sup\left\{\int_\Omega Y \, d\mathbb{P} \text{ }, Y \text{ is simple and } Y \leq X \text{ for } \omega \in \Omega\right\}$$

Integrals in Probability Space

3. If X is integrable , $\int_\Omega X \, d\mathbb{P} = \int_\Omega X^+ \, d\mathbb{P} - \int_\Omega X^- \, d\mathbb{P}$

4. Expectation over a region A, $\int_A X \, d\mathbb{P} = \int_\Omega \mathbb{I}_A \, X \, d\mathbb{P}$

5. Linearity , $\int_\Omega (\alpha X + \beta Y) d\mathbb{P} = \alpha \int_\Omega X d\mathbb{P} + \beta \int_\Omega Y d\mathbb{P}$

And Fatou's Lemma, Monotone Convergence and Dominated Convergence

Independence - Sets

Two sets $A \in \mathcal{F}$ and $B \in \mathcal{F}$ are independent if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A).\mathbb{P}(B)$$

If an outcome ω has happened. The $\mathbb{P}(A)$ basically means the unconditional probability of $\omega \in A$. Let's say we know that $\omega \in B$. Given this information, the updated probability of $\omega \in A$ is called the conditional probability of A given B and is denoted as $\mathbb{P}(A|B)$

The definition of conditional probability is $\mathbb{P}(A|B) = \dfrac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$

If A and B are independent,

the conditional probability of $\mathbb{P}(A|B)$ = unconditional probability $\mathbb{P}(A)$

Bayes' theorem is given as $\mathbb{P}(A|B) = \dfrac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}$

Independence - Sets

Whether 2 sets are independent, depends on the probability measure $\mathbb{P}$

Assume a random experiment of tossing a coin twice where the tosses are independent. We have $\mathbb{P}(H) = p$ and $\mathbb{P}(T) = q$, of course $p + q = 1$

Let's say $A = \{HH, HT\}$ = *"H on the first toss"* , $B = \{HT, TH\}$ = *"one H and one T"*

Let's calculate the probabilities

$$\mathbb{P}(A) = p^2 + pq = p$$

$$\mathbb{P}(B) = 2pq$$

$$\mathbb{P}(A)\mathbb{P}(B) = 2p^2q$$

$$\mathbb{P}(A \cap B) = \mathbb{P}(\{HT\}) = pq$$

Under what condition, A and B are independent ?

Independence - Sets

These sets are independent only if $p = q = \frac{1}{2}$

We can also check the conditional probability $\mathbb{P}(A|B) = \dfrac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \dfrac{pq}{2pq} = \dfrac{1}{2}$

So, for A and B to be independent, we need $\mathbb{P}(A) = \mathbb{P}(A|B) = \dfrac{1}{2}$ or $p = q = \frac{1}{2}$

We can use Bayes' Theorem to calculate $\mathbb{P}(B|A) = \dfrac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)} = \dfrac{\frac{1}{2}(2pq)}{p} = q$

which is also the unconditional probability of B i.e. $\mathbb{P}(B) = 2pq$ if $p = q = \frac{1}{2}$

Independence - σ −algebra

Let $\mathcal{G}$ and $\mathcal{H}$ be sub-sigma algebras of $\mathcal{F}$. We say $\mathcal{G}$ and $\mathcal{H}$ are independent if every set in $\mathcal{G}$ is independent of every set in $\mathcal{H}$ i.e., $\mathbb{P}(A \cap B) = \mathbb{P}(A).\mathbb{P}(B)$ for every $A \in \mathcal{G}$ and $B \in \mathcal{H}$

Let $\mathcal{G}$ be the σ −algebra containing the information of the first toss

$$\mathcal{G} = \{\varphi, \Omega, \{HH, HT\}, \{TH, TT\}\}$$

Let $\mathcal{H}$ be the σ −algebra containing the information of the second toss

$$\mathcal{H} = \{\varphi, \Omega, \{HH, TH\}, \{HT, TT\}\}$$

If we assume that the tosses are independent, then the σ −algebras will also be independent. This means any set from $\mathcal{G}$ and any set from $\mathcal{H}$ will be independent. For example,

$$\mathbb{P}(\{HH, HT\} \cap \{HH, TH\}) = \mathbb{P}(\{HH\}) = p^2$$

$$\mathbb{P}(\{HH, HT\}).\mathbb{P}(\{HH, TH\}) = (p^2 + pq)(p^2 + pq) = p^2$$

Independence - σ −algebra

Exercise – Show that under the following probability measure $\mathbb{P}$ for the random experiment of tossing a coin twice, the tosses are not independent.

$$\Omega = \{HH, HT, TH, TT\} \qquad \mathbb{P}\{HH\} = \frac{1}{9} \qquad \mathbb{P}\{HT\} = \frac{2}{9} \qquad \mathbb{P}\{TH\} = \frac{1}{3} \qquad \mathbb{P}\{TT\} = \frac{1}{3}$$

Quick sanity check if the $\mathbb{P}$ is legit. $\mathbb{P}(\omega) < 1$ for all $\omega \in \Omega$ and $\mathbb{P}(\Omega) = \frac{1}{9} + \frac{2}{9} + \frac{1}{3} + \frac{1}{3} = 1$

Let's derive the probabilities of $A =$ *"H on first toss"* and $B =$*"H on second toss"*

$$\mathbb{P}(A) = \mathbb{P}\{HH, HT\} = \frac{1}{9} + \frac{2}{9} = \frac{1}{3} \qquad \mathbb{P}(B) = \mathbb{P}\{HH, TH\} = \frac{1}{9} + \frac{1}{3} = \frac{4}{9}$$

Now $\mathbb{P}(A \cap B) = \mathbb{P}(\{HH, HT\} \cap \{HH, TH\}) = \mathbb{P}(\{HH\}) = \frac{1}{9} \neq \frac{4}{9}$

So, A and B are not independent

Independence – Random Variables

PEAKS2TAILS

Two random variables X and Y are independent if the sigma algebras generated by them $\sigma(X)$ and $\sigma(Y)$ are independent. That means any set A in $\sigma(X)$ will be independent of any set B in $\sigma(Y)$

We can define the independence of two random variables using induced measure as well.

The measure on any interval $A \subseteq \mathbb{R}$ induced by X is

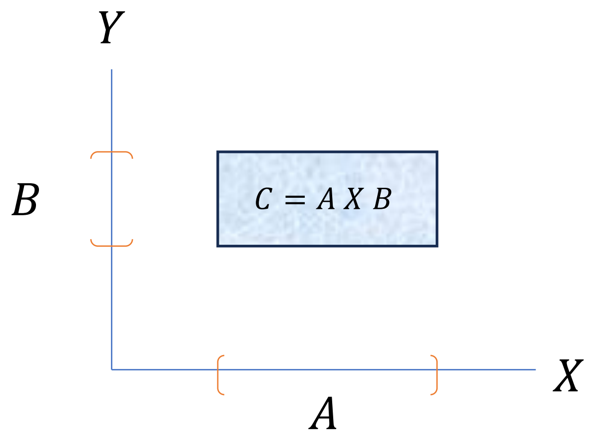
$$\mathcal{L}_X(A) = \mathbb{P}(X \in A)$$

The measure on any interval $B \subseteq \mathbb{R}$ induced by Y is

$$\mathcal{L}_Y(B) = \mathbb{P}(Y \in A)$$

The measure on any interval $C = A \text{ x } B \subseteq \mathbb{R}^2$ induced by the pair (X, Y) is

$$\mathcal{L}_{X,Y}(C) = \mathbb{P}(\{X \in A\} \cap \{Y \in B\})$$



Then X and Y are independent iff $\mathcal{L}_{X,Y}(A \text{ x } B) = \mathcal{L}_X(A)\ \mathcal{L}_Y(B)$

(Joint distribution is the product of marginal distributions)

Marginal & Joint distributions

For discrete random variables (X, Y), their joint distribution function $f_{X,Y}$ is the mass standing at (X, Y) in $\mathbb{R}^2$. Given the joint distribution function, we can derive the cumulative, marginal and conditional distributions.

$$\text{(Cumulative distribution)} := F_{X,Y}(x, y) = \mathbb{P}(X < x, Y < y) = \sum_{u=-\infty}^{u=x} \sum_{v=-\infty}^{v=y} f_{X,Y}(u, v)$$

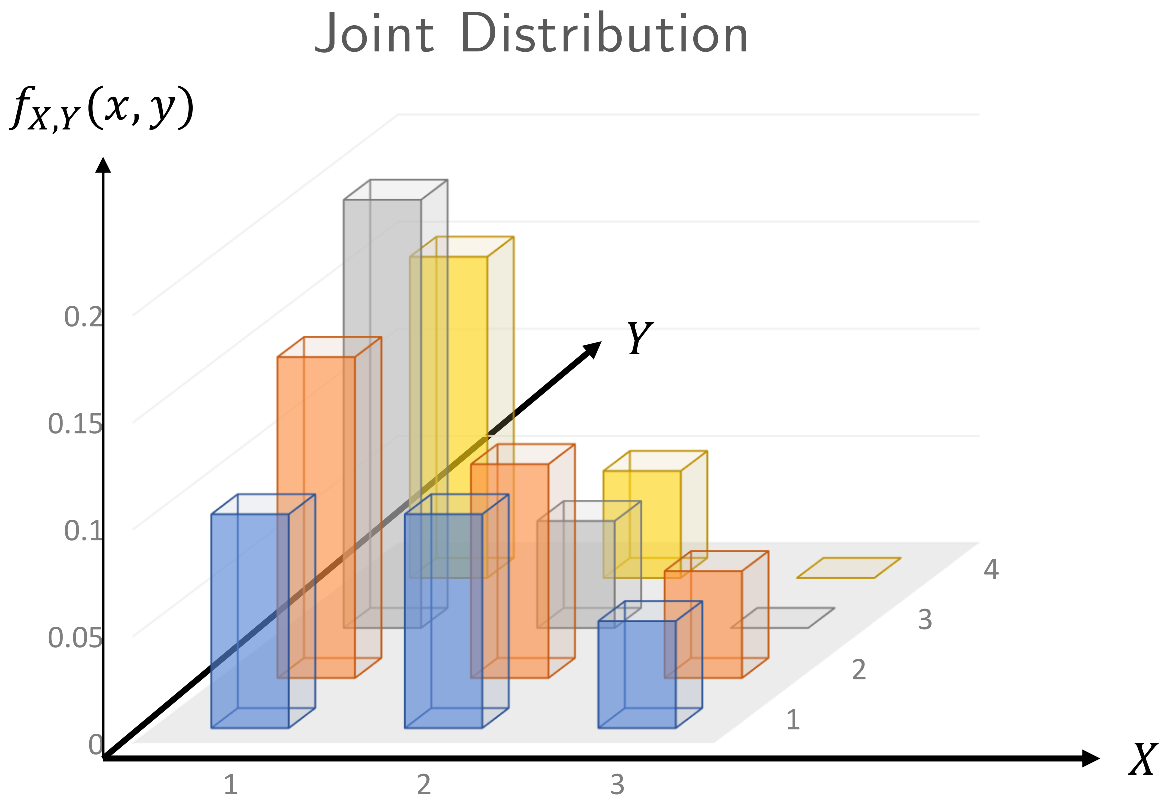
$$\text{(Marginal distribution of } X) := f_X(x) = \mathbb{P}(X = x) = \sum_{v=-\infty}^{v=\infty} f_{X,Y}(x, v)$$

$$\text{(Conditional distribution of } X|Y) := f_{X|Y}(x, y) = \mathbb{P}(X = x|Y = y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \qquad \text{(Joint / Marginal)}$$

Marginal & Joint distributions (Discrete)

Exercise 1 – Given random variables $X(\omega): \Omega \longrightarrow \{1,2,3\}$ and $Y(\omega): \Omega \longrightarrow \{1,2,3,4\}$ and their joint distribution $f_{X,Y}(x,y)$, derive the Marginal and Conditional distributions.

	x		
y	1	2	3
1	0.1	0.1	0.05
2	0.15	0.1	0.05
3	0.2	0.05	
4	0.15	0.05	

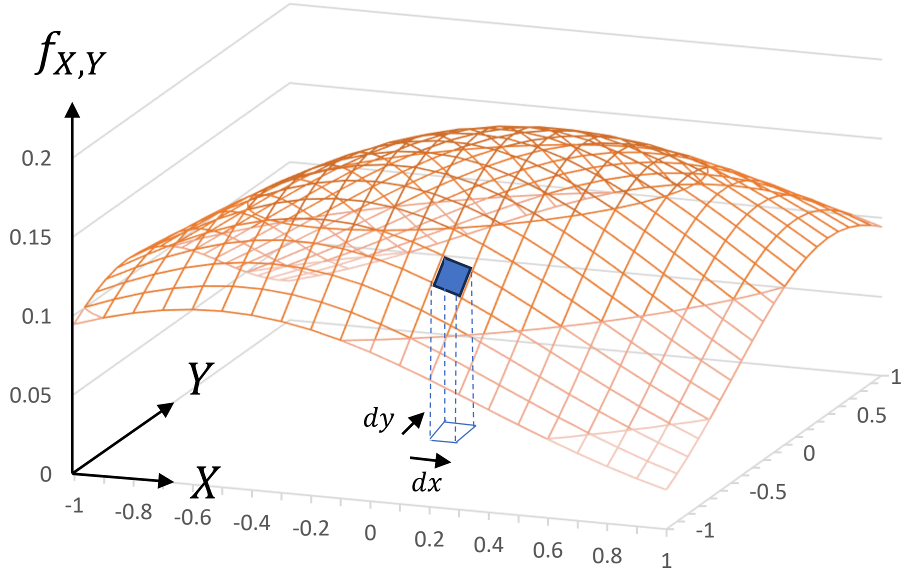


Marginal & Joint distributions (Discrete)

Exercise 2 – Given random variables $X(\omega): \Omega \longrightarrow \{1,2,3,4\}$ and $Y(\omega): \Omega \longrightarrow \{1,2,3\}$ and their joint distribution $f_{X,Y}(x, y) = \frac{x}{35 \text{ x } 2^{y-2}}$, derive the Marginal and Conditional distributions.

Marginal & Joint distributions (Continuous)

In case of continuous random variables, the summations get replaced by integration.



$$\mathbb{P}(x < X < x + dx, y < Y < y + dy) = f_{X,Y}dxdy$$

$$\mathbb{P}(x_1 < X < x_2, y_1 < Y < y_2) = \int_{x_1}^{x_2}\int_{y_1}^{y_2} f_{X,Y}(x,y)dxdy$$

$$F(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(x,y)dxdy$$ (Cumulative)

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx$$

(Marginal)

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

Marginal & Joint distributions (Continuous)

Exercise 3 – The bivariate joint density of (X, Y) is given as

$$f_{X,Y}(x, y) = \frac{2x + y}{c}\,, 10 \leq x \leq 20\,, -5 \leq y \leq 5$$

Determine the following.

(i) Value of c such that this density is a proper density

(ii) Find the probability $\mathbb{P}(10 \leq X \leq 15, Y \geq 0)$

(iii) Find the marginal density of X and evaluate $\mathbb{P}(12 \leq X \leq 14)$

(iv) Find the marginal density of Y and evaluate $\mathbb{P}(0 \leq Y \leq 5)$

(v) Find the conditional density $\mathbb{P}(X = x|Y = y)$. Are X, Y independent ?

Marginal & Joint distributions (Continuous)

Given the cumulative distribution function $F_{X,Y}(x,y)$, we can obtain the joint density

$$f_{X,Y}(x,y) = \frac{\partial^2 F}{\partial x \partial y}$$

Exercise 4 – Given the following cumulative distribution function

$$F_{X,Y}(x,y) = c(x^2 y + xy^2), 0 \leq x \leq 3\,, 0 \leq y \leq 4$$

Find

(i) the value of c ,

(ii) the joint density function

(iii) $\mathbb{P}(X \leq Y)$ and $\mathbb{P}(Y \leq X)$

Independence – Functions of Random Variables

Suppose random variables X and Y are independent. Consider two functions $g: \mathbb{R} \to \mathbb{R}$ and $h: \mathbb{R} \to \mathbb{R}$. Then $g(X)$ and $h(Y)$ are also independent random variables.

If $X_1, X_2, \ldots$ be a sequence of random variables. These random variables are independent if for every sequence of sets $A_1 \in \sigma(X_1)$, $A_2 \in \sigma(X_2)$, … and every $n \in \mathbb{N}$

$$\mathbb{P}(A_1 \cap A_2 \cap \cdots A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2) \ldots \mathbb{P}(A_n)$$

Correlation and Covariance

The Variance of a random variable X

$$\mathrm{Var}(X) = \mathbb{E}[X - \mathbb{E}[X]]^2$$

The Covariance of two random variables X and Y

$$\mathrm{Cov}\,(X,Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

Correlation is a standardized measure that lies between $(-1{,}1)$. It's given as

$$\rho\,(X,Y) = \frac{\mathrm{Cov}\,(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}$$

Properties of Expectation and Variance

Expectation is a linear operator

$$\mathbb{E}(\alpha X \pm \beta Y) = \alpha\mathbb{E}[X] \pm \beta\mathbb{E}[Y]$$

Variance is not a linear operator

$$\mathrm{Var}\,(\alpha X) = \alpha^2\mathrm{Var}\,(X)$$

$$\mathrm{Var}\,(X+Y) = \mathrm{Var}\,(X) + \mathrm{Var}\,(Y) + 2\mathrm{Cov}\,(X,Y)$$

Exercise 5 – Show that $\mathrm{Cov}\,(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$

Correlation and Independence

Does independence imply correlation = 0 ? (Yes)

Does correlation = 0 imply independence ? (No)

So, under independence, the variance of sum becomes the sum of variances. If $X_1, X_2, \ldots$ are a sequence of independent random variables, then

$$\mathrm{Var}\left(X_1 + X_2 + \cdots + X_n\right) = \mathrm{Var}\left(X_1\right) + \mathrm{Var}\left(X_2\right) + \cdots + \mathrm{Var}\left(X_n\right)$$

Law of Large numbers

If $X_1, X_2, \ldots$ are a sequence of independent and identically distributed (i.i.d) random variables, each with expectation μ and variance σ^2, then the sequence of averages

$$Y_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

converges to μ almost surely as $n \to \infty$

Exercise 6 – Prove the above result.

Central Limit Theorem

If $X_1, X_2, \ldots$ are a sequence of independent and identically distributed(i.i.d) random variables, each with expectation μ and variance σ^2, then the following quantity

$$Z_n = \frac{(X_1 - \mu) + (X_2 - \mu) + \cdots + (X_n - \mu)}{\sqrt{n}}$$

will always have an expectation μ and variance σ^2. But with $n \to \infty$, the distribution of Z_n approaches normal distribution. Another way of stating the same is

$$\lim_{n \to \infty} \sqrt{n}(Y_n - \mu) \xrightarrow{D} \mathcal{N}(0, \sigma^2)$$