Probability Primer

CS60077: Reinforcement Learning

Abir Das

IIT Kharagpur

Aug 13, 2021

Agenda

To brush up basics of probability and random variables.

Resources

- “Probability, Statistics, and Random Processes for Electrical Engineering”, 3rd Edition, Alberto Leon-Garcia - [PSRPEE] - Alberto Leon-Garcia
- “Machine Learning: A Probabilistic Perspective”, Kevin P. Murphy - [MLAPP] - Kevin Murphy:

Introduction

- Probability theory is the study of uncertainty.
- The mathematical treatise of probability is very sophisticated, and delves into a branch of analysis known as **measure theory**.
- We, however, will go through only basics of probability theory at a level appropriate for our Reinforcement Learning course.

Introduction

- Probability is the Mathematical language for quantifying *uncertainty*. The starting point is to specify random experiments, sample space and set of outcomes.
- A **random experiment** is an experiment in which the outcome varies in an unpredictable fashion when the experiment is repeated under the same conditions.
- An **outcome** is a result of the random experiment and it can not be decomposed in terms of other results. The **sample space** of a random experiment is defined as the set of all possible outcomes. An outcome and the sample space of a random experiment will be denoted as ζ and S respectively.

Introduction

- Probability is the Mathematical language for quantifying *uncertainty*. The starting point is to specify random experiments, sample space and set of outcomes.
- A **random experiment** is an experiment in which the outcome varies in an unpredictable fashion when the experiment is repeated under the same conditions.
- An **outcome** is a result of the random experiment and it can not be decomposed in terms of other results. The **sample space** of a random experiment is defined as the set of all possible outcomes. An outcome and the sample space of a random experiment will be denoted as ζ and S respectively.

Introduction

§ Examples of random experiment

- Flipping a coin
- Rolling a die
- Flipping a coin twice
- Pick a number X at random between zero and one, then pick a number Y at random between zero and X.

§ The corresponding sample spaces will be

- $S_1 = \{H, T\}$
- $S_2 = \{1, 2, 3, 4, 5, 6\}$
- $S_3 = \{HH, HT, TH, TT\}$
- $S_4 = \{(x, y) : 0 \leq y \leq x \leq 1\}$.

Introduction

§ Examples of random experiment

- Flipping a coin
- Rolling a die
- Flipping a coin twice
- Pick a number X at random between zero and one, then pick a number Y at random between zero and X.

§ The corresponding sample spaces will be

- $S_1 = \{H, T\}$
- $S_2 = \{1, 2, 3, 4, 5, 6\}$
- $S_3 = \{HH, HT, TH, TT\}$
- $S_4 = \{(x, y) : 0 \leq y \leq x \leq 1\}$.

Introduction

- Any subset E of the sample space S is known as an **event**. We, sometimes, are not interested in the occurrence of specific outcomes but rather in the occurrence of a combination of a few outcomes. This requires that we consider subsets of S
  - Getting even number when rolling a die, $E_2 = \{2, 4, 6\}$
  - Number of heads equal to number of tails when flipping a coin twice, $E_3 = \{HT, TH\}$
  - Two numbers differ by less than $1/10$, $E_4 = \{(x, y) : 0 \leq y \leq x \leq 1 \text{ and } |x - y| < 1/10\}$.
- We say that an event E occurs if the outcome ζ is in E

Introduction

- Any subset E of the sample space S is known as an **event**. We, sometimes, are not interested in the occurrence of specific outcomes but rather in the occurrence of a combination of a few outcomes. This requires that we consider subsets of S
  - Getting even number when rolling a die, $E_2 = \{2, 4, 6\}$
  - Number of heads equal to number of tails when flipping a coin twice, $E_3 = \{HT, TH\}$
  - Two numbers differ by less than $1/10$, $E_4 = \{(x, y) : 0 \leq y \leq x \leq 1 \text{ and } |x - y| < 1/10\}$.
- We say that an event E occurs if the outcome ζ is in E
- Three events are of special importance.
  - **Simple event** are the outcomes of random experiments.
  - **Sure event** is the sample space S which consists of all outcomes and hence always occurs.
  - **Impossible** or **null event** ϕ which contains no outcomes and hence never occurs.

Introduction

- Any subset E of the sample space S is known as an **event**. We, sometimes, are not interested in the occurrence of specific outcomes but rather in the occurrence of a combination of a few outcomes. This requires that we consider subsets of S
  - Getting even number when rolling a die, $E_2 = \{2, 4, 6\}$
  - Number of heads equal to number of tails when flipping a coin twice, $E_3 = \{HT, TH\}$
  - Two numbers differ by less than $1/10$, $E_4 = \{(x, y) : 0 \leq y \leq x \leq 1 \text{ and } |x - y| < 1/10\}$.
- We say that an event E occurs if the outcome ζ is in E
- Three events are of special importance.
  - **Simple event** are the outcomes of random experiments.
  - **Sure event** is the sample space S which consists of all outcomes and hence always occurs.
  - **Impossible** or **null event** ϕ which contains no outcomes and hence never occurs.

Introduction

- Any subset E of the sample space S is known as an **event**. We, sometimes, are not interested in the occurrence of specific outcomes but rather in the occurrence of a combination of a few outcomes. This requires that we consider subsets of S
  - Getting even number when rolling a die, $E_2 = \{2, 4, 6\}$
  - Number of heads equal to number of tails when flipping a coin twice, $E_3 = \{HT, TH\}$
  - Two numbers differ by less than $1/10$, $E_4 = \{(x, y) : 0 \leq y \leq x \leq 1 \text{ and } |x - y| < 1/10\}$.
- We say that an event E occurs if the outcome ζ is in E
- Three events are of special importance.
  - **Simple event** are the outcomes of random experiments.
  - **Sure event** is the sample space S which consists of all outcomes and hence always occurs.
  - **Impossible** or **null event** ϕ which contains no outcomes and hence never occurs.

Introduction

- **Set of events** (or **event space**) $\mathcal{F}$: A set whose elements are subsets of the sample space (*i.e.*, events). $\mathcal{F} = \{A : A \subseteq S\}$. $\mathcal{F}$ is really a "set of sets".
- $\mathcal{F}$ should satisfy the following three properties.
  - $\phi \in \mathcal{F}$
  - $A \in \mathcal{F} \implies A^c (\triangleq S \setminus A) \in \mathcal{F}$
  - $A_1, A_2, \cdots \in \mathcal{F} \implies \cup_i A_i \in \mathcal{F}$

Introduction

- Probabilities are numbers assigned to events of $\mathcal{F}$ that indicate how "likely" it is that the events will occur when a random experiment is performed.
- Let a random experiment has sample space S and event space $\mathcal{F}$. Probability of an event A is a function $P : \mathcal{F} \rightarrow \mathbb{R}$ that satisfies the following properties
  - $P(A) \geq 0, \ \forall A \in \mathcal{F}$
  - $P(S) = 1$
  - If $A_1, A_2, \cdots \in \mathcal{F}$ are disjoint events ($i.e.$, $A_i \cap A_j = \phi$ for $i \neq j$) then, $P(\cup_i A_i) = \sum_i P(A_i)$
- These three properties are called the **Axioms of Probability**.

Introduction

§ Probabilities are numbers assigned to events of $\mathcal{F}$ that indicate how "likely" it is that the events will occur when a random experiment is performed.

§ Let a random experiment has sample space S and event space $\mathcal{F}$. Probability of an event A is a function $P : \mathcal{F} \rightarrow \mathbb{R}$ that satisfies the following properties

- $P(A) \geq 0, \ \forall A \in \mathcal{F}$
- $P(S) = 1$
- If $A_1, A_2, \cdots \in \mathcal{F}$ are disjoint events (*i.e.*, $A_i \cap A_j = \phi$ for $i \neq j$) then, $P(\cup_i A_i) = \sum_i P(A_i)$

§ These three properties are called the **Axioms of Probability**.

Introduction

- Probabilities are numbers assigned to events of $\mathcal{F}$ that indicate how "likely" it is that the events will occur when a random experiment is performed.
- Let a random experiment has sample space S and event space $\mathcal{F}$. Probability of an event A is a function $P : \mathcal{F} \rightarrow \mathbb{R}$ that satisfies the following properties
  - $P(A) \geq 0, \ \forall A \in \mathcal{F}$
  - $P(S) = 1$
  - If $A_1, A_2, \cdots \in \mathcal{F}$ are disjoint events (*i.e.*, $A_i \cap A_j = \phi$ for $i \neq j$) then, $P(\cup_i A_i) = \sum_i P(A_i)$
- These three properties are called the **Axioms of Probability**.

Introduction

§ Properties

- $P(A^c) = 1 - P(A)$
- $P(A) \leq 1$
- $P(\phi) = 0$
- If $A \subseteq B$, then $P(A) \leq P(B)$.
- $P(A \cap B) \leq \min(P(A), P(B))$
- $P(A \cup B) \leq P(A) + P(B)$

Conditional Probability

- **Conditional probability** provides whether two events are related in the sense that knowledge about the occurrence of one, say B, alters the likelihood of occurrence of the other say, A.
- This conditional probability is defined as,

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

- Two events A and B are **independent** (denoted as $A \perp B$) if the knowledge of occurrence of one does not change the likelihood of occurrence of the other. This translates to the condition for independence as,

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

Conditional Probability

- **Conditional probability** provides whether two events are related in the sense that knowledge about the occurrence of one, say B, alters the likelihood of occurrence of the other say, A.
- This conditional probability is defined as,

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

- Two events A and B are **independent** (denoted as $A \perp B$) if the knowledge of occurrence of one does not change the likelihood of occurrence of the other. This translates to the condition for independence as,

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

Total Probability Theorem

§ Let $B_1, B_2, \cdots, B_n$ be exhaustive and mutually exclusive events such that each of these events has positive probabilities. Then for any event A, the *total probability theorem* says,

$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i) \tag{1}$$

§ **Proof:** Since, $B_1, B_2, \cdots, B_n$ are exhaustive (*i.e.*, their union covers the whole sample space), $A = (A \cap B_1) \cup (A \cap B_2) \cup \cdots (A \cap B_n)$

$$\begin{aligned} P(A) &= P((A \cap B_1) \cup (A \cap B_2) \cup \cdots (A \cap B_n)) \\ &= P(A \cap B_1) + P(A \cap B_2) + \cdots + P(A \cap B_n) \\ &\text{(as } B_i\text{'s are mutually exclusive)} \\ &= \sum_{i=1}^{n} P(A \cap B_i) = \sum_{i=1}^{n} P(A|B_i)P(B_i) \end{aligned}$$

Total Probability Theorem

- Let $B_1, B_2, \cdots, B_n$ be exhaustive and mutually exclusive events such that each of these events has positive probabilities. Then for any event A, the *total probability theorem* says,

$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i) \quad (1)$$

- **Proof:** Since, $B_1, B_2, \cdots, B_n$ are exhaustive (*i.e.*, their union covers the whole sample space), $A = (A \cap B_1) \cup (A \cap B_2) \cup \cdots (A \cap B_n)$

$$P(A) = P\big((A \cap B_1) \cup (A \cap B_2) \cup \cdots (A \cap B_n)\big)$$
$$= P(A \cap B_1) + P(A \cap B_2) + \cdots + P(A \cap B_n)$$

(as B_i's are mutually exclusive)

$$= \sum_{i=1}^{n} P(A \cap B_i) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$$

Total Probability Theorem

§ Let $B_1, B_2, \cdots, B_n$ be exhaustive and mutually exclusive events such that each of these events has positive probabilities. Then for any event A, the *total probability theorem* says,

$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i) \tag{1}$$

§ **Proof:** Since, $B_1, B_2, \cdots, B_n$ are exhaustive (*i.e.*, their union covers the whole sample space), $A = (A \cap B_1) \cup (A \cap B_2) \cup \cdots (A \cap B_n)$

$$\begin{aligned} P(A) &= P\big((A \cap B_1) \cup (A \cap B_2) \cup \cdots (A \cap B_n)\big) \\ &= P(A \cap B_1) + P(A \cap B_2) + \cdots + P(A \cap B_n) \\ &\text{(as } B_i\text{'s are mutually exclusive)} \\ &= \sum_{i=1}^{n} P(A \cap B_i) = \sum_{i=1}^{n} P(A|B_i)P(B_i) \end{aligned}$$

Total Probability Theorem

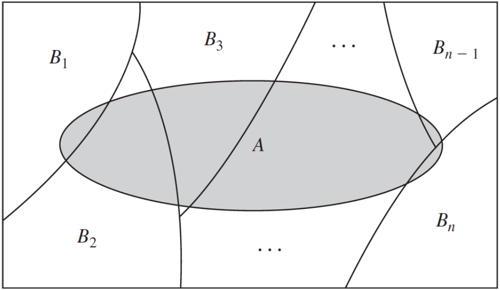


Figure credit: *[PSRPEE] - Alberto Leon-Garcia*

- This is also known as **marginalization** operation.
- Such exhaustive and mutually exclusive events $B_1, B_2, \cdots, B_n$ are also said to form a **partition** of the sample space.

Bayes Rule

- The total probability theorem is often used in conjunction with the Bayes' Rule that relates conditional probabilities of the form $P(B|A)$ with conditional probabilities of the form $P(A|B)$.
- Let the events $B_1, B_2, \cdots, B_n$ partitions a sample space such that each of the $P(B_i)$'s are non-negative. The Bayes' rule states,

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A)} = \frac{P(A|B_i)P(B_i)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)} \quad (2)$$

- Bayes' rule is a very important tool for inference in machine learning. A can be thought of as the "effect" and B_i's are several "causes" that can result in the effect. From the probabilities of the causes (B_i's) resulting in the effect (A) and the probability of the causes (B_i's) to occur frequently, the probability that a particular cause (B_i) is the reason behind the effect (A) is computed.

Random Variables

- Statistics and Machine Learning are concerned with data. The link to sample space and events to data is **Random Variables**.
- A random variable is a mapping ($X : S \to \mathbb{R}$) from the sample space to real values that assigns a real number ($X(\zeta)$) to each outcome (ζ) in the sample space of a random experiment.

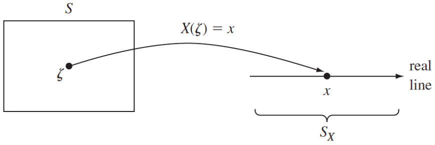


FIGURE 3.1
A random variable assigns a number $X(\zeta)$ to each outcome ζ in the sample space S of a random experiment.

Figure credit: *[PSRPEE] - Alberto Leon-Garcia*

- We will use the following notation: capital letters denote random variables, *e.g.*, X or Y, and lower case letters denote possible values of the random variables, *e.g.*, x or y.

Random Variables

§ An example from [PSRPEE] - Alberto Leon-Garcia

Example 3.1 Coin Tosses

A coin is tossed three times and the sequence of heads and tails is noted. The sample space for this experiment is $S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$. Let X be the number of heads in the three tosses. X assigns each outcome ζ in S a number from the set $S_X = \{0, 1, 2, 3\}$. The table below lists the eight outcomes of S and the corresponding values of X.

ζ:	HHH	HHT	HTH	THH	HTT	THT	TTH	TTT
$X(\zeta)$:	3	2	2	2	1	1	1	0

X is then a random variable taking on values in the set $S_X = \{0, 1, 2, 3\}$.

§ Since the value of a random variable is determined by the outcome of the experiment, we may assign probabilities to the possible values of the random variable.

$$P(X = x) = P(\{\zeta \in S; X(\zeta) = x\}) \quad (3)$$

Random Variables

§ An example from [PSRPEE] - Alberto Leon-Garcia

Example 3.1 Coin Tosses

A coin is tossed three times and the sequence of heads and tails is noted. The sample space for this experiment is $S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$. Let X be the number of heads in the three tosses. X assigns each outcome ζ in S a number from the set $S_X = \{0, 1, 2, 3\}$. The table below lists the eight outcomes of S and the corresponding values of X.

ζ:	HHH	HHT	HTH	THH	HTT	THT	TTH	TTT
$X(\zeta)$:	3	2	2	2	1	1	1	0

X is then a random variable taking on values in the set $S_X = \{0, 1, 2, 3\}$.

§ Since the value of a random variable is determined by the outcome of the experiment, we may assign probabilities to the possible values of the random variable.

$$P(X = x) = P(\{\zeta \in S;\ X(\zeta) = x\}) \quad (3)$$

Random Variables

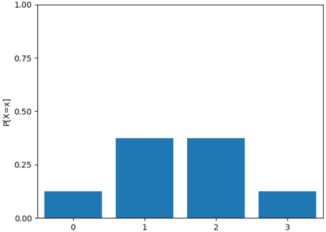
ζ:	HHH	HHT	HTH	THH	HTT	THT	TTH	TTT
$X(\zeta)$:	3	2	2	2	1	1	1	0

$$P[X=0] = P[\{TTT\}] = \frac{1}{8}$$

$$P[X=1] = P[\{HTT, THT, TTH\}] = P[\{HTT\}] + P[\{THT\}] + P[\{TTH\}] = \frac{3}{8}$$

$$P[X=2] = P[\{HHT, HTH, THH\}] = P[\{HHT\}] + P[\{HTH\}] + P[\{THH\}] = \frac{3}{8}$$

$$P[X=3] = P[\{HHH\}] = \frac{1}{8}$$



Plot generated by: discrete_prob_dist_plot

from [MLAPP] - Kevin Murphy

Discrete Random Variables and PMF

- A **discrete random variable** X is defined as a random variable that can take at most a countable number of possible values, *i.e.*, $S_X = \{x_1, x_2, x_3, \cdots\}$.
- A discrete random variable is said to be **finite** if its range is finite, *i.e.*, $S_X = \{x_1, x_2, x_3, \cdots, x_n\}$.
- The probabilities of events involving a discrete random variable X forms the **Probability Mass Function (PMF)** of X and it is defined as (ref eqn. (3)),

$$P_X(x) = P(X = x) = P(\{\zeta \in S; X(\zeta) = x\} \text{ for real } x) \qquad (4)$$

Discrete Random Variables and PMF

- A **discrete random variable** X is defined as a random variable that can take at most a countable number of possible values, *i.e.*, $S_X = \{x_1, x_2, x_3, \cdots\}$.
- A discrete random variable is said to be **finite** if its range is finite, *i.e.*, $S_X = \{x_1, x_2, x_3, \cdots, x_n\}$.
- The probabilities of events involving a discrete random variable X forms the **Probability Mass Function (PMF)** of X and it is defined as (ref eqn. (3)),

$$P_X(x) = P(X = x) = P(\{\zeta \in S; X(\zeta) = x\} \text{ for real } x) \quad (4)$$

- Note that $P_X(x)$ is a function of x over the real line, and that $P_X(x)$ can be nonzero only at the values $\{x_1, x_2, x_3, \cdots\}$

Discrete Random Variables and PMF

- A **discrete random variable** X is defined as a random variable that can take at most a countable number of possible values, *i.e.*, $S_X = \{x_1, x_2, x_3, \cdots\}$.
- A discrete random variable is said to be **finite** if its range is finite, *i.e.*, $S_X = \{x_1, x_2, x_3, \cdots, x_n\}$.
- The probabilities of events involving a discrete random variable X forms the **Probability Mass Function (PMF)** of X and it is defined as (ref eqn. (3)),

$$P_X(x) = P(X = x) = P(\{\zeta \in S;\ X(\zeta) = x\} \text{ for real } x) \tag{4}$$

- Note that $P_X(x)$ is a function of x over the real line, and that $P_X(x)$ can be nonzero only at the values $\{x_1, x_2, x_3, \cdots\}$

Discrete Random Variables and PMF

- § A **discrete random variable** X is defined as a random variable that can take at most a countable number of possible values, *i.e.*, $S_X = \{x_1, x_2, x_3, \cdots\}$.
- § A discrete random variable is said to be **finite** if its range is finite, *i.e.*, $S_X = \{x_1, x_2, x_3, \cdots, x_n\}$.
- § The probabilities of events involving a discrete random variable X forms the **Probability Mass Function (PMF)** of X and it is defined as (ref eqn. (3)),

$$P_X(x) = P(X = x) = P(\{\zeta \in S;\, X(\zeta) = x\} \text{ for real } x) \quad (4)$$

- § Note that $P_X(x)$ is a function of x over the real line, and that $P_X(x)$ can be nonzero only at the values $\{x_1, x_2, x_3, \cdots\}$

Continuous Random Variables and PDF

- Random variables with a continuous range of possible experimental values are quite common.
- X is a **continuous random variable** if there exists a non-negative function $f_X(x)$, defined for all real $x \in (-\infty, \infty)$, having the property that for any set B of real numbers, $P(X \in B) = \int_B f_X(x)dx$. The function $f_X(x)$ is called the **probability density function (PDF)** of the random variable X.
- Some properties of PDFs
  - $P(-\infty < X < \infty) = \int_{-\infty}^{\infty} f_X(x)dx = 1$
  - $P(a \le X \le b) = \int_a^b f_X(x)dx$
  - If we let $a = b$ in the preceding, then $P(X = a) = \int_a^a f_X(x)dx = 0$
  - This means $P(a \le X \le b) = P(a < X < b) = P(a \le X < b) = P(a < X \le b)$

Continuous Random Variables and PDF

- Random variables with a continuous range of possible experimental values are quite common.
- X is a **continuous random variable** if there exists a non-negative function $f_X(x)$, defined for all real $x \in (-\infty, \infty)$, having the property that for any set B of real numbers, $P(X \in B) = \int_B f_X(x)dx$. The function $f_X(x)$ is called the **probability density function (PDF)** of the random variable X.
- Some properties of PDFs
  - $P(-\infty < X < \infty) = \int_{-\infty}^{\infty} f_X(x)dx = 1$
  - $P(a \leq X \leq b) = \int_a^b f_X(x)dx$
  - If we let $a = b$ in the preceding, then $P(X = a) = \int_a^a f_X(x)dx = 0$
  - This means $P(a \leq X \leq b) = P(a < X < b) = P(a \leq X < b) = P(a < X \leq b)$

Continuous Random Variables and PDF

- Random variables with a continuous range of possible experimental values are quite common.
- X is a **continuous random variable** if there exists a non-negative function $f_X(x)$, defined for all real $x \in (-\infty, \infty)$, having the property that for any set B of real numbers, $P(X \in B) = \int_B f_X(x)dx$. The function $f_X(x)$ is called the **probability density function (PDF)** of the random variable X.
- Some properties of PDFs
  - $P(-\infty < X < \infty) = \int_{-\infty}^{\infty} f_X(x)dx = 1$
  - $P(a \leq X \leq b) = \int_a^b f_X(x)dx$
  - If we let $a = b$ in the preceding, then $P(X = a) = \int_a^a f_X(x)dx = 0$
  - This means $P(a \leq X \leq b) = P(a < X < b) = P(a \leq X < b) = P(a < X \leq b)$

Cumulative Distribution Function

- We have defined PMF and PDF for discrete and continuous random variables respectively.
- **Cumulative Distribution Function (CDF)** is a concept that is applicable to both discrete and continuous random variables. It is defined as,

$$F_X(x) = P(X \leq x) = \begin{cases} \sum\limits_{k \leq x} P_X(k) & X: \text{ discrete} \\ \int\limits_{-\infty}^{x} f_X(t)dt & X: \text{ continuous} \end{cases} \quad (5)$$

- For continuous random variables, the cumulative distribution function $F_X(x)$ is differentiable everywhere. Naturally, in these cases, PDF is the derivative of the CDF.

$$f_X(x) = \frac{dF_X(x)}{dx}$$

Cumulative Distribution Function

- We have defined PMF and PDF for discrete and continuous random variables respectively.
- **Cumulative Distribution Function (CDF)** is a concept that is applicable to both discrete and continuous random variables. It is defined as,

$$F_X(x) = P(X \leq x) = \begin{cases} \sum\limits_{k \leq x} P_X(k) & X: \text{ discrete} \\ \int\limits_{-\infty}^{x} f_X(t)dt & X: \text{ continuous} \end{cases} \quad (5)$$

- For continuous random variables, the cumulative distribution function $F_X(x)$ is differentiable everywhere. Naturally, in these cases, PDF is the derivative of the CDF.

$$f_X(x) = \frac{dF_X(x)}{dx}$$

Cumulative Distribution Function

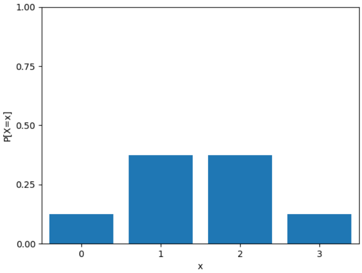
- We have defined PMF and PDF for discrete and continuous random variables respectively.
- **Cumulative Distribution Function (CDF)** is a concept that is applicable to both discrete and continuous random variables. It is defined as,

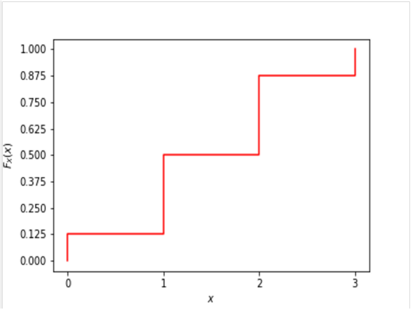
$$F_X(x) = P(X \le x) = \begin{cases} \sum\limits_{k \le x} P_X(k) & X: \text{ discrete} \\ \int\limits_{-\infty}^{x} f_X(t)dt & X: \text{ continuous} \end{cases} \tag{5}$$

- For continuous random variables, the cumulative distribution function $F_X(x)$ is differentiable everywhere. Naturally, in these cases, PDF is the derivative of the CDF.

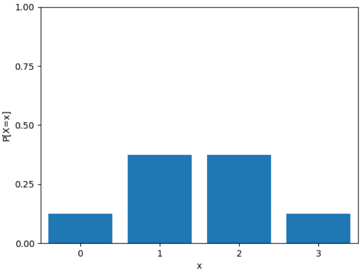
$$f_X(x) = \frac{dF_X(x)}{dx}$$

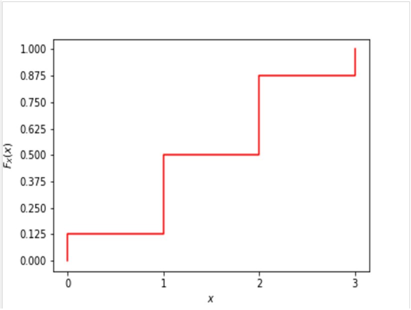
Cumulative Distribution Function





Cumulative Distribution Function





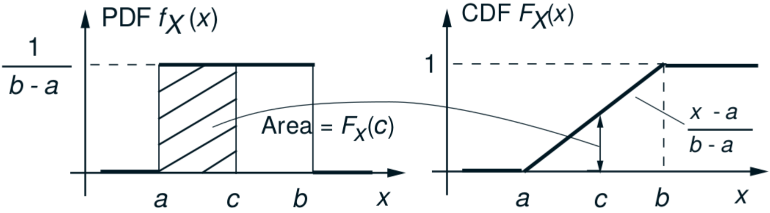


Fig credit: MIT Course: 6.041-6.43, Lecture Notes

Expectation

- The **expected value/expectation/mean** of a random variable is defined as:

$$\mathbb{E}[X] = \begin{cases} \sum_x x P_X(x) & \text{when } X \text{ is discrete} \\ \int x f_X(x) dx & \text{when } X \text{ is continuous} \end{cases} \quad (6)$$

- **Functions of random variable**: If $Y = g(X)$ is a function of a random variable X, then Y is also a random variable, since it provides a numerical value for each possible outcome.

- For a function of the random variable $Y = g(X)$, the expectation is, similarly, defined as,

$$\mathbb{E}[g(X)] = \begin{cases} \sum_x g(x) P_X(x) & \text{when } X \text{ is discrete} \\ \int g(x) f_X(x) dx & \text{when } X \text{ is continuous} \end{cases} \quad (7)$$

Expectation

- The **expected value/expectation/mean** of a random variable is defined as:

$$\mathbb{E}[X] = \begin{cases} \sum_x x P_X(x) & \text{when } X \text{ is discrete} \\ \int x f_X(x) dx & \text{when } X \text{ is continuous} \end{cases} \quad (6)$$

- **Functions of random variable**: If $Y = g(X)$ is a function of a random variable X, then Y is also a random variable, since it provides a numerical value for each possible outcome.

- For a function of the random variable $Y = g(X)$, the expectation is, similarly, defined as,

$$\mathbb{E}[g(X)] = \begin{cases} \sum_x g(x) P_X(x) & \text{when } X \text{ is discrete} \\ \int g(x) f_X(x) dx & \text{when } X \text{ is continuous} \end{cases} \quad (7)$$

Expectation

- The **expected value/expectation/mean** of a random variable is defined as:

$$\mathbb{E}[X] = \begin{cases} \sum_x x P_X(x) & \text{when } X \text{ is discrete} \\ \int x f_X(x) dx & \text{when } X \text{ is continuous} \end{cases} \tag{6}$$

- **Functions of random variable**: If $Y = g(X)$ is a function of a random variable X, then Y is also a random variable, since it provides a numerical value for each possible outcome.
- For a function of the random variable $Y = g(X)$, the expectation is, similarly, defined as,

$$\mathbb{E}[g(X)] = \begin{cases} \sum_x g(x) P_X(x) & \text{when } X \text{ is discrete} \\ \int g(x) f_X(x) dx & \text{when } X \text{ is continuous} \end{cases} \tag{7}$$

Variance

- § $\mathbb{E}[X]$ is also referred to as the **first moment** of X. Similarly the second moment is defined as $\mathbb{E}[X^2]$ and in general, the n^{th} moment as $\mathbb{E}[X^n]$
- § Another quantity of interest is the variance of a random variable x, denoted as $\mathrm{var}(X)$ and defined as $\mathbb{E}\left[\left(X - \mathbb{E}[X]\right)^2\right]$. Variance provides a measure of dispersion of X around its mean $\mathbb{E}[X]$.
- § Another measure of dispersion is the standard deviation of X, which is defined as the square root of the variance $\sigma_X = \sqrt{\mathrm{var}(X)}$

Variance

- $\mathbb{E}[X]$ is also referred to as the **first moment** of X. Similarly the second moment is defined as $\mathbb{E}[X^2]$ and in general, the n^{th} moment as $\mathbb{E}[X^n]$
- Another quantity of interest is the variance of a random variable x, denoted as $\text{var}(X)$ and defined as $\mathbb{E}\left[\left(X - \mathbb{E}[X]\right)^2\right]$. Variance provides a measure of dispersion of X around its mean $\mathbb{E}[X]$.
- Another measure of dispersion is the standard deviation of X, which is defined as the square root of the variance $\sigma_X = \sqrt{\text{var}(X)}$
- Note that, using the rule for expected value of functions of random variables variance can be computed as,

$$\text{var}(X) = \mathbb{E}\left[\left(X - \mathbb{E}[X]\right)^2\right] = \begin{cases} \sum_x \left(X - \mathbb{E}[X]\right)^2 P_X(x) & \text{for discrete } X \\ \int \left(X - \mathbb{E}[X]\right)^2 f_X(x)dx & \text{for continuous } X \end{cases} \quad (8)$$

Variance

- $\mathbb{E}[X]$ is also referred to as the **first moment** of X. Similarly the second moment is defined as $\mathbb{E}[X^2]$ and in general, the n^{th} moment as $\mathbb{E}[X^n]$
- Another quantity of interest is the variance of a random variable x, denoted as $\text{var}(X)$ and defined as $\mathbb{E}\left[\left(X - \mathbb{E}[X]\right)^2\right]$. Variance provides a measure of dispersion of X around its mean $\mathbb{E}[X]$.
- Another measure of dispersion is the standard deviation of X, which is defined as the square root of the variance $\sigma_X = \sqrt{\text{var}(X)}$
- Note that, using the rule for expected value of functions of random variables variance can be computed as,

$$\text{var}(X) = \mathbb{E}\left[\left(X - \mathbb{E}[X]\right)^2\right] = \begin{cases} \sum_x \left(X - \mathbb{E}[X]\right)^2 P_X(x) & \text{for discrete } X \\ \int \left(X - \mathbb{E}[X]\right)^2 f_X(x)dx & \text{for continuous } X \end{cases} \tag{8}$$

Variance

- $\mathbb{E}[X]$ is also referred to as the **first moment** of X. Similarly the second moment is defined as $\mathbb{E}[X^2]$ and in general, the n^{th} moment as $\mathbb{E}[X^n]$
- Another quantity of interest is the variance of a random variable x, denoted as $\mathrm{var}(X)$ and defined as $\mathbb{E}\left[\left(X - \mathbb{E}[X]\right)^2\right]$. Variance provides a measure of dispersion of X around its mean $\mathbb{E}[X]$.
- Another measure of dispersion is the standard deviation of X, which is defined as the square root of the variance $\sigma_X = \sqrt{\mathrm{var}(X)}$
- Note that, using the rule for expected value of functions of random variables variance can be computed as,

$$\mathrm{var}(X) = \mathbb{E}\left[\left(X - \mathbb{E}[X]\right)^2\right] = \begin{cases} \sum_x \left(X - \mathbb{E}[X]\right)^2 P_X(x) & \text{for discrete } X \\ \int \left(X - \mathbb{E}[X]\right)^2 f_X(x)dx & \text{for continuous } X \end{cases} \tag{8}$$

Properties

§ Expectation

- $\mathbb{E}[a] = a$ for any constant $a \in \mathbb{R}$
- $\mathbb{E}[af(X)] = a\mathbb{E}[f(X)]$ for any constant $a \in \mathbb{R}$
- $\mathbb{E}[f(X) + g(X)] = \mathbb{E}[f(X)] + \mathbb{E}[g(X)]$

§ Variance

- $\text{var}(X) = \mathbb{E}\left[\left(X - \mathbb{E}[X]\right)^2\right] = \mathbb{E}[X^2] - \left[\mathbb{E}[X]\right]^2$
- $\text{var}(a) = 0$ for any constant $a \in \mathbb{R}$
- $\text{var}(af(X)) = a^2\,\text{var}(f(X))$ for any constant $a \in \mathbb{R}$

Some Common Random Variables

Discrete Random Variables

- **Bernoulli** random variable: Takes two values 1 and 0 (or 'Head' and 'Tail'). The PMF is given by,

$$P_X(x) = \begin{cases} p & \text{if } x = 1 \\ 1-p & \text{if } x = 0 \end{cases} \tag{9}$$

  This is also written as $P_X(x) = p^x(1-p)^{1-x}$
- It is used to model situations with just two random outcomes *e.g.*, tossing a coin once.
- For $X \sim \text{Ber}(p), \mathbb{E}(X) = p$ and $\text{var}(X) = p(1-p)$.

Some Common Random Variables

Discrete Random Variables

§ **Binomial** random variable: is used to model more complex situation *e.g.*, the number of heads if a coin is tossed n times. The PMF is given by,

$$P_X(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \cdots, n. \quad (10)$$

§ For $X \sim \text{Bin}(n, p), \mathbb{E}(X) = np$ and $\text{var}(X) = np(1-p)$.

Some Common Random Variables

Discrete Random Variables

- **Poisson** random variable: models situations where the events occur completely at random in time or space. The random variable counts the number of occurrences of the event in a certain time period or in a certain region in space. The PMF is given by,

$$P_X(x) = P(X = x) = \frac{\lambda^x}{x!}e^{-\lambda}, \quad x = 0, 1, 2, \cdots \tag{11}$$

  where λ is the average number of occurrences of the event in that specified time interval or region in space.
- For $X \sim \text{Poisson}(\lambda), \mathbb{E}(X) = \lambda$ and $\text{var}(X) = \lambda$.

Some Common Random Variables

Continuous Random Variables

- **Uniform** random variable: X is a uniform random variable on the interval (a, b) if its probability density function is given by,

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a \leq x \leq b \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

- For $X \sim \text{Uniform}(a, b), \mathbb{E}(X) = \frac{a+b}{2}$ and $\text{var}(X) = \frac{(b-a)^2}{12}$.

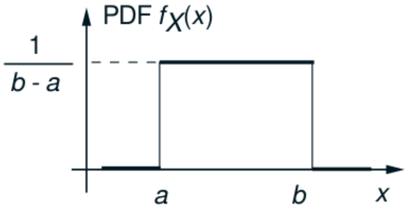


Fig credit: MIT Course: 6.041-6.43, Lecture Notes

Some Common Random Variables

Continuous Random Variables

- **Exponential** random variable: X is a exponential random variable if its probability density function is given by,

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

- For $X \sim \text{Exponential}(\lambda), \mathbb{E}(X) = \frac{1}{\lambda}$ and $\text{var}(X) = \frac{1}{\lambda^2}$.

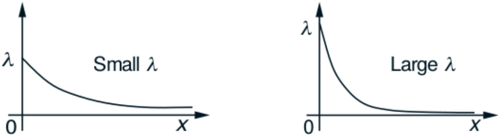


Fig credit: MIT Course: 6.041-6.43, Lecture Notes

Some Common Random Variables

Continuous Random Variables

- **Gaussian/Normal** random variable: X is a Gaussian/Normal random variable if its probability density function is given by,

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{14}$$

- For $X \sim \text{Gaussian}(\mu, \sigma^2), \mathbb{E}(X) = \mu$ and $\text{var}(X) = \sigma^2$.
- Gaussianity is Preserved by Linear Transformations. If $X \sim \text{Gaussian}(\mu, \sigma^2)$ and if a, b are scalars, the random variable $Y = aX + b$ is also Gaussian with mean and variance $\mathbb{E}(X) = a\mu + b$ and $\text{var}(X) = a^2\sigma^2$ respectively.

Two Random Variables

§ Many random experiments involve several random variables. For example, temperature and pressure of a room during different days.

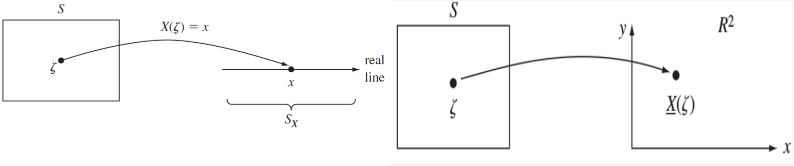


Figure credit: *[PSRPEE] - Alberto Leon-Garcia*

§ Consider two discrete random variables X and Y associated with the same experiment. We will use the notation $P(X = x, Y = y)$ to denote $P(X = x \text{ and } Y = y)$.

Two Random Variables

- The **Joint PMF** of the two random variables X and Y is defined as,

$$\begin{aligned} P_{X,Y}(x,y) &= P(X=x, Y=y) \\ &= P\big(\{\zeta \in S;\ X(\zeta)=x, Y(\zeta)=y\} \text{ for real } x \text{ and } y\big) \end{aligned} \tag{15}$$

- $P_X(x)$ and $P_Y(y)$ are sometimes referred to as the **marginal PMFs**, to distinguish them from the joint PMF.
- The marginal and the joint PMFs are related in the following way (ref eqn. (1), the total probability theorem),

$$P_X(x) = \sum_y P_{X,Y}(x,y) \text{ and } P_Y(y) = \sum_x P_{X,Y}(x,y) \tag{16}$$

Two Random Variables

- Similar to PDFs for single random variable, **joint PDF** for two continuous random variables is defined. for sets A and B of real numbers,

$$P(X \in A, Y \in B) = \int_B \int_A f_{X,Y}(x,y)dxdy \tag{17}$$

- Similarly, **joint CDF** is also defined.

$$F_{X,Y}(x,y) = P(X \leq x, Y \leq y) = \begin{cases} \sum\limits_{l \leq y} \sum\limits_{k \leq x} P_{X,Y}(k,l) & X,Y: \text{ discrete} \\ \int\limits_{-\infty}^{y} \int\limits_{-\infty}^{x} f_{X,Y}(u,v)dudv & X,Y: \text{ continuous} \end{cases} \tag{18}$$

- Differentiation for continuous random variables, yields

$$f_{X,Y}(x,y) = \frac{dF_{X,Y}(x,y)}{dydx}$$

Some Useful Relations

- Marginal CDF can be obtained by setting the value of the other Random Variable to ∞, *i.e.*, $F_X(x) = F_{X,Y}(x, \infty)$ and $F_Y(y) = F_{X,Y}(\infty, y)$.
- Similar relations exist between marginal and joint PDFs. $f_X(x) = \int\limits_{-\infty}^{\infty} f_{X,Y}(x, y)dy$ and $f_Y(y) = \int\limits_{-\infty}^{\infty} f_{X,Y}(x, y)dx$
- Conditional PMF and Marginal PMF for discrete variables are related as, $P_{Y|X}(y|x) = \frac{P_{X,Y}(x,y)}{P_X(x)}$ assuming that $P_X(x) \neq 0$.
- Similar relation is there for continuous random variables. $f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$ provided $f_X(x) \neq 0$.

Joint Expectations

- Similar expectation and moment rules exist for joint moments and expectation as in the case of a single random variable.
- Considering $Z = g(X, Y)$ as a function of two random variables, the expectation of Z can be found as,

$$\mathbb{E}[Z] = \begin{cases} \int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} g(x, y) f_{X,Y}(x, y) dx dy & X, Y \text{continuous} \\ \sum\limits_{i} \sum\limits_{j} g(x_i, y_j) P_{X,Y}(x_i, y_n) & X, Y \text{discrete} \end{cases} \tag{19}$$

- Expectation of a sum of random variables is the sum of the expectations of the random variables.

$$\mathbb{E}[X_1 + X_2 + X_3 + \cdots] = \mathbb{E}[X_1] + \mathbb{E}[X_2] + \mathbb{E}[X_3] + \cdots \tag{20}$$

Joint Moments, Correlation, and Covariance

§ The jk^{th} **joint moment** of X and Y is defined as,

$$\mathbb{E}[X^jY^k] = \begin{cases} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^jy^k f_{X,Y}(x,y)dxdy & X,Y \text{continuous} \\ \sum_m \sum_n x_m^j y_n^k P_{X,Y}(x_m,y_n) & X,Y \text{discrete} \end{cases} \quad (21)$$

§ When $j = k = 1$, the corresponding moment $\mathbb{E}[XY]$ gives the correlation between X and Y. If $\mathbb{E}[XY] = 0$, X and Y are said to be **orthogonal**.

§ The jk^{th} **central moment** of X and Y is defined as $\mathbb{E}\left[(X-\mathbb{E}(X))^j(Y-\mathbb{E}(Y))^k\right]$

Joint Moments, Correlation, and Covariance

§ The jk^{th} **joint moment** of X and Y is defined as,

$$\mathbb{E}[X^jY^k] = \begin{cases} \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} x^j y^k f_{X,Y}(x,y)dxdy & X, Y \text{continuous} \\ \sum\limits_m \sum\limits_n x_m^j y_n^k P_{X,Y}(x_m, y_n) & X, Y \text{discrete} \end{cases} \tag{21}$$

§ When $j = k = 1$, the corresponding moment $\mathbb{E}[XY]$ gives the correlation between X and Y. If $\mathbb{E}[XY] = 0$, X and Y are said to be **orthogonal**.

§ The jk^{th} **central moment** of X and Y is defined as $\mathbb{E}\Big[\big(X - \mathbb{E}(X)\big)^j \big(Y - \mathbb{E}(Y)\big)^k\Big]$

§ When $j = k = 1$, the corresponding central moment $\mathbb{E}\big[\big(X - \mathbb{E}(X)\big)\big(Y - \mathbb{E}(Y)\big)\big]$ is called the **covariance** between X and Y.

Joint Moments, Correlation, and Covariance

§ The jk^{th} **joint moment** of X and Y is defined as,

$$\mathbb{E}[X^jY^k] = \begin{cases} \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} x^jy^kf_{X,Y}(x,y)dxdy & X,Y \text{continuous} \\ \sum\limits_{m}\sum\limits_{n} x_m^jy_n^kP_{X,Y}(x_m,y_n) & X,Y \text{discrete} \end{cases} \tag{21}$$

§ When $j=k=1$, the corresponding moment $\mathbb{E}[XY]$ gives the correlation between X and Y. If $\mathbb{E}[XY]=0$, X and Y are said to be **orthogonal**.

§ The jk^{th} **central moment** of X and Y is defined as $\mathbb{E}\Big[\big(X-\mathbb{E}(X)\big)^j\big(Y-\mathbb{E}(Y)\big)^k\Big]$

§ When $j=k=1$, the corresponding central moment $\mathbb{E}\big[\big(X-\mathbb{E}(X)\big)\big(Y-\mathbb{E}(Y)\big)\big]$ is called the **covariance** between X and Y.

Joint Moments, Correlation, and Covariance

- The jk^{th} **joint moment** of X and Y is defined as,

$$\mathbb{E}[X^jY^k] = \begin{cases} \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} x^j y^k f_{X,Y}(x,y)dxdy & X, Y \text{continuous} \\ \sum\limits_m \sum\limits_n x_m^j y_n^k P_{X,Y}(x_m, y_n) & X, Y \text{discrete} \end{cases} \tag{21}$$

- When $j = k = 1$, the corresponding moment $\mathbb{E}[XY]$ gives the correlation between X and Y. If $\mathbb{E}[XY] = 0$, X and Y are said to be **orthogonal**.
- The jk^{th} **central moment** of X and Y is defined as $\mathbb{E}\Big[\big(X - \mathbb{E}(X)\big)^j \big(Y - \mathbb{E}(Y)\big)^k\Big]$
- When $j = k = 1$, the corresponding central moment $\mathbb{E}\big[\big(X - \mathbb{E}(X)\big)\big(Y - \mathbb{E}(Y)\big)\big]$ is called the **covariance** between X and Y.

Joint Moments, Correlation, and Covariance

- Covariance can also be expressed as $\text{COV}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$
- If X and Y are independent, then $\text{COV}(X,Y) = 0$, *i.e.*, $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$
- Correlation coefficient turns covariance into a normalized scale between -1 to 1.

$$\rho_{X,Y} = \frac{\text{COV}(X,Y)}{\sqrt{\text{VAR}(X)}\sqrt{\text{VAR}(Y)}} = \frac{\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]}{\sqrt{\text{VAR}(X)}\sqrt{\text{VAR}(Y)}} \tag{22}$$

Joint Moments, Correlation, and Covariance

- Covariance can also be expressed as $\text{COV}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$
- If X and Y are independent, then $\text{COV}(X, Y) = 0$, *i.e.*, $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$
- Correlation coefficient turns covariance into a normalized scale between -1 to 1.

$$\rho_{X,Y} = \frac{\text{COV}(X, Y)}{\sqrt{\text{VAR}(X)}\sqrt{\text{VAR}(Y)}} = \frac{\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]}{\sqrt{\text{VAR}(X)}\sqrt{\text{VAR}(Y)}} \quad (22)$$

- $\rho_{X,Y} = 0$ means X and Y are uncorrelated. Then $\text{COV}(X, Y) = 0$.

Joint Moments, Correlation, and Covariance

- Covariance can also be expressed as $\text{COV}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$
- If X and Y are independent, then $\text{COV}(X,Y) = 0$, *i.e.*, $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$
- Correlation coefficient turns covariance into a normalized scale between -1 to 1.

$$\rho_{X,Y} = \frac{\text{COV}(X,Y)}{\sqrt{\text{VAR}(X)}\sqrt{\text{VAR}(Y)}} = \frac{\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]}{\sqrt{\text{VAR}(X)}\sqrt{\text{VAR}(Y)}} \quad (22)$$

- $\rho_{X,Y} = 0$ means X and Y are uncorrelated. Then $\text{COV}(X,Y) = 0$.
- If X and Y are independent, then they are uncorrelated, but the reverse is not always true (true always for Gaussian random variables). Check Section 5.6.2 of [PSRPEE] for more details and examples.

Joint Moments, Correlation, and Covariance

- Covariance can also be expressed as COV$(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$
- If X and Y are independent, then COV$(X, Y) = 0$, *i.e.*, $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$
- Correlation coefficient turns covariance into a normalized scale between -1 to 1.

$$\rho_{X,Y} = \frac{\text{COV}(X,Y)}{\sqrt{\text{VAR}(X)}\sqrt{\text{VAR}(Y)}} = \frac{\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]}{\sqrt{\text{VAR}(X)}\sqrt{\text{VAR}(Y)}} \tag{22}$$

- $\rho_{X,Y} = 0$ means X and Y are uncorrelated. Then COV$(X, Y) = 0$.
- If X and Y are independent, then they are uncorrelated, but the reverse is not always true (true always for Gaussian random variables). Check Section 5.6.2 of [PSRPEE] for more details and examples.

Joint Moments, Correlation, and Covariance

- Covariance can also be expressed as $\text{COV}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$
- If X and Y are independent, then $\text{COV}(X,Y) = 0$, *i.e.*, $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$
- Correlation coefficient turns covariance into a normalized scale between -1 to 1.

$$\rho_{X,Y} = \frac{\text{COV}(X,Y)}{\sqrt{\text{VAR}(X)}\sqrt{\text{VAR}(Y)}} = \frac{\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]}{\sqrt{\text{VAR}(X)}\sqrt{\text{VAR}(Y)}} \tag{22}$$

- $\rho_{X,Y} = 0$ means X and Y are uncorrelated. Then $\text{COV}(X,Y) = 0$.
- If X and Y are independent, then they are uncorrelated, but the reverse is not always true (true always for Gaussian random variables). Check Section 5.6.2 of [PSRPEE] for more details and examples.

Conditional Expectation

- The conditional expectation of Y given $X = x$ is defined as,

$$\mathbb{E}[Y|x] = \int_{-\infty}^{\infty} y f_{Y|x}(y|x) dy \tag{23}$$

- The conditional expectation $\mathbb{E}(Y|x)$ can be viewed as defining a function of x, $g(x) = \mathbb{E}(Y|x)$. As x, is a result of a random experiment, $\mathbb{E}(Y|x)$ is a random variable. So, we can find its expectation as,

$$\mathbb{E}\left[\mathbb{E}[Y|x]\right] = \int_{-\infty}^{\infty} \mathbb{E}[Y|x] f_X(x) dx = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y f_{Y|x}(y|x) f_X(x) dx dy \tag{24}$$

- With some simple manipulation of the double integral it can be easily shown that $\mathbb{E}[Y] = \mathbb{E}\left[\mathbb{E}[Y|x]\right]$. Sometimes, to remove confusion it is also written as $\mathbb{E}_Y[Y] = \mathbb{E}_X\left[\mathbb{E}_Y[Y|x]\right]$ where the subscripts of the expectation sign denotes the expection w.r.t. that random variable.

Conditional Expectation

§ The conditional expectation of Y given $X = x$ is defined as,

$$\mathbb{E}[Y|x] = \int_{-\infty}^{\infty} y f_{Y|x}(y|x) dy \tag{23}$$

§ The conditional expectation $\mathbb{E}(Y|x)$ can be viewed as defining a function of x, $g(x) = \mathbb{E}(Y|x)$. As x, is a result of a random experiment, $\mathbb{E}(Y|x)$ is a random variable. So, we can find its expectation as,

$$\mathbb{E}\big[\mathbb{E}[Y|x]\big] = \int_{-\infty}^{\infty} \mathbb{E}[Y|x] f_X(x) dx = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y f_{Y|x}(y|x) f_X(x) dx dy \tag{24}$$

§ With some simple manipulation of the double integral it can be easily shown that $\mathbb{E}[Y] = \mathbb{E}\big[\mathbb{E}[Y|x]\big]$. Sometimes, to remove confusion it is also written as $\mathbb{E}_Y[Y] = \mathbb{E}_X\big[\mathbb{E}_Y[Y|x]\big]$ where the subscripts of the expectation sign denotes the expection w.r.t. that random variable.

Conditional Expectation

§ The conditional expectation of Y given $X = x$ is defined as,

$$\mathbb{E}[Y|x] = \int_{-\infty}^{\infty} y f_{Y|x}(y|x)dy \tag{23}$$

§ The conditional expectation $\mathbb{E}(Y|x)$ can be viewed as defining a function of x, $g(x) = \mathbb{E}(Y|x)$. As x, is a result of a random experiment, $\mathbb{E}(Y|x)$ is a random variable. So, we can find its expectation as,

$$\mathbb{E}\left[\mathbb{E}[Y|x]\right] = \int_{-\infty}^{\infty} \mathbb{E}[Y|x] f_X(x)dx = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y f_{Y|x}(y|x) f_X(x)dxdy \tag{24}$$

§ With some simple manipulation of the double integral it can be easily shown that $\mathbb{E}[Y] = \mathbb{E}\left[\mathbb{E}[Y|x]\right]$. Sometimes, to remove confusion it is also written as $\mathbb{E}_Y[Y] = \mathbb{E}_X\left[\mathbb{E}_Y[Y|x]\right]$ where the subscripts of the expectation sign denotes the expection w.r.t. that random variable.

Conditional Independence

- X and Y are **conditionally independent** given Z iff the conditional joint can be written as product of conditional marginals,

$$X \perp\!\!\!\perp Y|Z \Leftrightarrow P(X,Y|Z) = P(X|Z)P(Y|Z) \tag{25}$$

- Conditional independence also implies,

$$X \perp\!\!\!\perp Y|Z \Rightarrow P(X|Y,Z) = P(X|Z) \text{ and } P(Y|X,Z) = P(Y|Z) \tag{26}$$

- Z causes X and Y. Given it is 'raining', we don't need to know whether 'frogs are out' to predict if 'ground is wet'.

Conditional Independence

- X and Y are **conditionally independent** given Z iff the conditional joint can be written as product of conditional marginals,

$$X \perp\!\!\!\perp Y|Z \Leftrightarrow P(X,Y|Z) = P(X|Z)P(Y|Z) \tag{25}$$

- Conditional independence also implies,

$$X \perp\!\!\!\perp Y|Z \Rightarrow P(X|Y,Z) = P(X|Z) \text{ and } P(Y|X,Z) = P(Y|Z) \tag{26}$$

- Z causes X and Y. Given it is 'raining', we don't need to know whether 'frogs are out' to predict if 'ground is wet'.

Conditional Independence

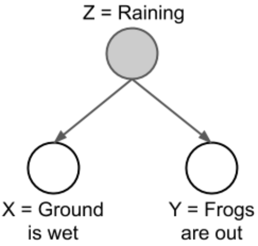
- X and Y are **conditionally independent** given Z iff the conditional joint can be written as product of conditional marginals,

$$X \perp\!\!\!\perp Y | Z \Leftrightarrow P(X, Y|Z) = P(X|Z)P(Y|Z) \tag{25}$$

- Conditional independence also implies,

$$X \perp\!\!\!\perp Y | Z \Rightarrow P(X|Y, Z) = P(X|Z) \text{ and } P(Y|X, Z) = P(Y|Z) \tag{26}$$

- Z causes X and Y. Given it is 'raining', we don't need to know whether 'frogs are out' to predict if 'ground is wet'.



Multiple Random Variables

- The notions and ideas can be generalized to more than two random variables. A **vector random variable** $\mathbf{X}$ is a function that assigns a vector of real numbers to each outcome ζ in the sample space S of a random experiment.
- Uppercase boldface letters are generally used to denote vector random variables. By convention, it is a column vector. Each X_i can be thought of as a random variable itself.

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} = \left[X_1, X_2, \cdots, X_n\right]^T$$

- Possible values of the vector random variable are denoted by $\mathbf{x} = \left[x_1, x_2, \cdots, x_n\right]^T$

Multiple Random Variables

§ The **Joint PMF** of n-dimensional discrete random vector $\mathbf{X}$

$$P_{\mathbf{X}}(\mathbf{x}) = P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n) \tag{27}$$

§ Relation between the marginal and the joint PMFs,

$$P_{X_1}(x_1) = \sum_{x_2} \cdots \sum_{x_n} P_{\mathbf{X}}(\mathbf{x}) \tag{28}$$

§ Similarly, **joint CDF** is also defined.

$$F_{\mathbf{X}}(\mathbf{x}) = P(X_1 \leq x_1, X_2 \leq x_2, \cdots, X_n \leq x_n)$$
$$= \begin{cases} \sum\limits_{j \leq x_1} \sum\limits_{k \leq x_2} \cdots \sum\limits_{l \leq x_n} P_{\mathbf{X}}([x_1, x_2, \cdots, x_n]^T) & \mathbf{X}: \text{ discrete} \\ \int\limits_{-\infty}^{x_1} \int\limits_{-\infty}^{x_2} \cdots \int\limits_{-\infty}^{x_n} f_{\mathbf{X}}([u, v, \cdots, w]^T) du dv \cdots dw & \mathbf{X}: \text{ continuous} \end{cases} \tag{29}$$

Multiple Random Variables

§ The **joint PDF** of n-dimensional continuous random vector $\mathbf{X}$

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{\partial^n F_{\mathbf{X}}(\mathbf{x})}{\partial x_1 \partial x_2 \cdots \partial x_3} \tag{30}$$

§ The **marginal PDF**

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbf{X}}([x_1, x_2, x_3, \cdots, x_n]^T)\, dx_2 dx_3 \cdots dx_n \tag{31}$$

§ The **conditional PDF**

$$f_{X_1/X_2,\cdots,X_n}(x_1/x_2, \cdots, x_n) = \frac{f_{\mathbf{X}}(\mathbf{x})}{f_{X_2,\cdots,X_n}(x_2, \cdots, x_n)} \tag{32}$$

§ **Chain rule**

$$\begin{aligned} f(x_1, x_2, \cdots, x_n) &= f(x_n|x_1, \cdots, x_{n-1}) f(x_1, \cdots, x_{n-1}) \\ &= f(x_n|x_1, \cdots, x_{n-1}) f(x_{n-1}|x_1, \cdots, x_{n-2}) f(x_1, \cdots, x_{n-2}) \\ &= f(x_1) \prod_{i=2}^{n} f(x_i|x_1, x_2, \cdots, x_{i-1}) \end{aligned}$$

Multiple Random Variables

- There's also natural generalization of **independence**.

$$f(x_1, x_2, \cdots, x_n) = f(x_1)f(x_2)\cdots f(x_n) \quad (34)$$

- **Expectation**: Consider an arbitrary function $g : \mathbb{R}^n \rightarrow \mathbb{R}$. The expected value is,

$$\mathbb{E}[g(\mathbf{X})] = \int_{\mathbb{R}^n} g(\mathbf{X}) f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x} \quad (35)$$

- If g is a function from $\mathbb{R}^n$ to $\mathbb{R}^m$, then the expected value of g is the element-wise expected values of the output vector, *i.e.*, if $\mathbf{g}(\mathbf{x}) = [g_1(\mathbf{x}), g_2(\mathbf{x}, \cdots, g_n(\mathbf{x}))]^T$, then $\mathbb{E}[\mathbf{g}(\mathbf{x})] = \left[\mathbb{E}[g_1(\mathbf{x})], \mathbb{E}[g_2(\mathbf{x}], \cdots, \mathbb{E}[g_n(\mathbf{x}))]\right]^T$

Multiple Random Variables

§ There's also natural generalization of **independence**.

$$f(x_1, x_2, \cdots, x_n) = f(x_1)f(x_2)\cdots f(x_n) \tag{34}$$

§ **Expectation**: Consider an arbitrary function $g : \mathbb{R}^n \rightarrow \mathbb{R}$. The expected value is,

$$\mathbb{E}\big[g(\mathbf{X})\big] = \int\limits_{\mathbb{R}^n} g(\mathbf{X}) f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x} \tag{35}$$

§ If g is a function from $\mathbb{R}^n$ to $\mathbb{R}^m$, then the expected value of g is the element-wise expected values of the output vector, *i.e.*, if $\mathbf{g}(\mathbf{x}) = \big[g_1(\mathbf{x}), g_2(\mathbf{x}, \cdots, g_n(\mathbf{x}))\big]^T$, then
$\mathbb{E}\big[\mathbf{g}(\mathbf{x})\big] = \Big[\mathbb{E}\big[g_1(\mathbf{x})\big], \mathbb{E}\big[g_2(\mathbf{x}\big], \cdots, \mathbb{E}\big[g_n(\mathbf{x}))\big]\Big]^T$

Multiple Random Variables

§ **Covariance matrix**: For a random vector $\mathbf{X}\in\mathbb{R}^n$, covariance matrix $\mathbf{\Sigma}$ is $n\times n$ square matrix whose entries are given by $\mathbf{\Sigma}_{ij}=\text{Cov}(X_i,X_j)$.

$$\mathbf{\Sigma} = \begin{bmatrix} \text{Var}(X_1,X_1) & \text{Cov}(X_1,X_2) & \cdots & \text{Cov}(X_1,X_n) \\ \text{Cov}(X_2,X_1) & \text{Var}(X_2,X_2) & \cdots & \text{Var}(X_2,X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_n,X_1) & \text{Cov}(X_n,X_2) & \cdots & \text{Var}(X_n,X_n) \end{bmatrix}$$

$$= \begin{bmatrix} \mathbb{E}[X_1^2]-\mathbb{E}[X_1]\mathbb{E}[X_1] & \cdots & \mathbb{E}[X_1X_n]-\mathbb{E}[X_1]\mathbb{E}[X_n] \\ \mathbb{E}[X_2X_1]-\mathbb{E}[X_2]\mathbb{E}[X_1] & \cdots & \mathbb{E}[X_2X_n]-\mathbb{E}[X_2]\mathbb{E}[X_n] \\ \vdots & \ddots & \vdots \\ \mathbb{E}[X_nX_1]-\mathbb{E}[X_n]\mathbb{E}[X_1] & \cdots & \mathbb{E}[X_n^2]-\mathbb{E}[X_n]\mathbb{E}[X_n] \end{bmatrix}$$

$$= \begin{bmatrix} \mathbb{E}[X_1^2] & \cdots & \mathbb{E}[X_1X_n] \\ \mathbb{E}[X_2X_1] & \cdots & \mathbb{E}[X_2X_n] \\ \vdots & \ddots & \vdots \\ \mathbb{E}[X_nX_1] & \cdots & \mathbb{E}[X_n^2] \end{bmatrix} - \begin{bmatrix} \mathbb{E}[X_1]\mathbb{E}[X_1] & \cdots & \mathbb{E}[X_1]\mathbb{E}[X_n] \\ \mathbb{E}[X_2]\mathbb{E}[X_1] & \cdots & \mathbb{E}[X_2]\mathbb{E}[X_n] \\ \vdots & \ddots & \vdots \\ \mathbb{E}[X_n]\mathbb{E}[X_1] & \cdots & \mathbb{E}[X_n]\mathbb{E}[X_n] \end{bmatrix}$$

$$= \mathbb{E}[\mathbf{X}\mathbf{X}^T] - \mathbb{E}[\mathbf{X}]\mathbb{E}[\mathbf{X}^T] = \mathbb{E}\left[(\mathbf{X}-\mathbb{E}[\mathbf{X}])(\mathbf{X}-\mathbb{E}[\mathbf{X}])^T\right]$$

(36)