# Introduction to Probability and Statistics

Study Note

Han Cheol Moon
School of Computer Science and Engineering
Nanyang Technological University
Singapore
hancheol001@e.ntu.edu.sg

June 30, 2024

# Contents

# Part I

# Introduction to Probability

# Chapter 1

# Introduction to Probability

## 1.1 Introduction

The probability of event $A$ as

$$P(A) = \frac{\text{Number of times A occurs}}{\text{Total number of outcomes}}$$

This commonsense understanding of probability is called the *relative frequency definition*.

# Chapter 2

# Combinatorics

## 2.1 Multiplication Principle

Suppose that we perform $r$ experiments such that the $k$-th experiment has $n_k$ possible outcomes, for $k = 1, 2, \ldots, r$. Then there are a total of $n_1 \times n_2 \times n_3 \times \cdots \times n_r$ possible outcomes for the sequence of r experiments.

## 2.2 Ordered Sampling with Replacement

Here we have a set with $n$ elements (*e.g.,* $A = \{1, 2, 3, \ldots n\}$), and we want to draw $k$ samples from the set such that ordering matters and repetition is allowed. For example, if $A = \{1, 2, 3\}$ and $k = 2$, there are 9 different possibilities. In general, we can argue that there are$k$ positions in the chosen list: (Position 1, Position 2, ..., Position $k$). There are $n$ options for each position. Thus, when ordering matters and repetition is allowed, the total number of ways to choose $k$ objects from a set with $n$ elements is

$$n \times n \times \cdots \times n = n^k.$$

## 2.3 Ordered Sampling without Replacement: Permutations

Consider the same setting as above, but now repetition is not allowed. For example, if $A = \{1, 2, 3\}$ and $k = 2$, there are 6 different possibilities. In general, we can argue that there are $k$ positions in the chosen list: (Position 1, Position 2, ..., Position $k$). There are $n$ options for the first position, $(n-1)$ options for the second position (since one element has already been allocated to the first position and cannot be chosen here), $(n-2)$ options for the third position, and $(n-k+1)$ options for the $k$-th position. Thus, when ordering matters and repetition is not allowed, the total number of ways to choose $k$ objects from a set with n elements is

$$n \times (n-1) \times \cdots \times (n-k+1).$$

It si called a $k$ permutation of the elements in set $A$. We use the following notation:

$$P_k^n = n \times (n-1) \times \cdots \times (n-k+1).$$

Note that if $k$ is larger than $n$, then $P_k^n = 0$.

**Example:** Birthday problem or birthday paradox is a problem that If $k$ people are at a party, what is the probability that at least two of them have the same birthday? Suppose that there are $n = 365$ days in a year and all days are equally likely to be the birthday of a specific person.

$$P(A) = 1 - \frac{P_k^n}{n^k}.$$

The reason this is called a paradox is that $P(A)$ is numerically different from what most people expect. For example, if there are $k = 23$ people in the party, what do you guess is the probability that at least two of them have the same birthday, $P(A)$? The answer is .5073, which is much higher than what most people guess. The probability crosses 99 percent when the number of peoples reaches 57. But why is the probability higher than what we expect?

It is important to note that in the birthday problem, neither of the two people are chosen beforehand. To better answer this question, let us look at a different problem: I am in a party with $k-1$ people. What is the probability that at least one person in the party has the same birthday as mine? Well, we need to choose the birthdays of $k-1$ people, the total number of ways to do this is $n^{k-1}$. The total number of ways to choose the birthdays so that no one has my birthday is $(n-1)^{k-1}$. Thus, the probability that at least one person has the same birthday as mine is

$$P(B) = 1 - \left(\frac{n-1}{n}\right)^{k-1}.$$

Now, if $k = 23$, this probability is only $P(B) = 0.0586$, which is much smaller than the corresponding $P(A) = 0.5073$. The reason is that event $B$ is looking only at the case where one person in the party has the same birthday as me. This is a much smaller event than event $A$ which looks at all possible pairs of people. Thus, $P(A)$ is much larger than $P(B)$. We might guess that the value of $P(A)$ is much lower than it actually is, because we might confuse it with $P(B)$.

**Permutations of $n$ elements:** An $n$-permutation of $n$ elements is just called a permutation of those elements. In this case $k = n$ and we have

$$\begin{aligned} P_n^n &= n \times (n-1) \times \cdots \times (n-n+1) \\ &= n \times (n-1) \times \cdots \times 1, \end{aligned}$$

which is denoted $n!$. We can rewrite as

$$P_k^n = \frac{n!}{(n-k)!}.$$

## 2.4 Unordered Sampling without Replacement: Combinations

Here we have a set with $n$ elements, *e.g.,* $A = \{1, 2, 3, \ldots, n\}$ and we want to draw $k$ samples from the set such that ordering does not matter and repetition is not allowed. Thus, we basically want to choose a $k$-element subset of $A$, which we also call a $k$-combination of the set $A$. For example if $A = \{1, 2, 3\}$ and $k = 2$, there are 3 different possibilities. We show the number of $k$-element subsets of $A$ by

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

This is also called the *binomial coefficient.* This is because the coefficients in the binomial theorem are given by

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}.$$

An intuitive way to understand this is that there are $n \times (n-1) \times \cdots \times (n-k+1)$ ways to place items and the $k \times \cdots \times 1$ ways to order the times, which can be ignored.

A simple way to find $\binom{n}{k}$ is to compare it with $P_k^n$. Note that the difference between the two is ordering.

$$P_k^n = \binom{n}{k} \times k!.$$

**Example 1:** I choose 3 cards from the standard deck of cards. What is the probability that these cards contain at least one ace?

- The sample space contains all possible ways to choose 3 cards from 52 cards.
- There are $52 - 4 = 48$ non-ace cards

**Example 2:** How many distinct sequences can we make using 3 letter "A"s and 5 letter "B"s? (AAABBBBB, AABABBBB, .)

You can think of this problem in the following way. You have 3+5=8 positions to fill with letters A or B. From these 8 positions, you need to choose 3 of them for "A"s. Whatever is left will be filled with "B"s. Thus the total number of ways is

$$\binom{8}{3}.$$

Equivalently, you chould have chosen the locations for Bs.

$$\binom{8}{5}.$$

The same argument can be repeated for general $n$ and $k$ to conclude

$$\binom{n}{k} = \binom{n}{n-k}.$$

## 2.5 Bernoulli Trials and Binomial Distribution

A *Bernoulli Trial* is a random experiment that has two possible outcomes which we can label as "success" and "failure", such as

- You toss a coin. The possible outcomes are H and T.

We usually denote the probability of success by $p$ and probability of failure by $q = 1 - p$. If we have an experiment in which we perform $n$ independent Bernoulli trials and count the total number of successes, we call it a binomial experiment. For example, you may toss a coin $n$ times repeatedly and be interested in the total number of heads.

**Example:** Suppose that I have a coin for which $P(H) = p$ and $P(T) = 1 - p$. I toss the coin 5 times.

- $P(THHHH) = p(T) \times p(H) \cdots = (1-p)p^4$
- $P(HTHHH) = (1-p)p^4$
- $P(HHTHH) = (1-p)p^4$
- $B = \{THHHH, HTHHH, HHTHH, HHHTH, HHHHT\}$, $P(B) = 5p^4(1-p)$
- Let $C = \{TTHHH, THTHH, \ldots\}$.

$$\begin{aligned} P(C) &= P(TTHHH) + P(THTHH) + \ldots \\ &= |C|p^3(1-p)^2 \end{aligned}$$

- The $|C|$ is the total number of distinct sequences that you can create using two tails and three heads.
$$\binom{5}{3}.$$
- Therefore,
$$P(C) = \binom{5}{3}p^3(1-p)^2$$

Now we can define *Binomial Formula*: For $n$ independent Bernoulli trials where each trial has success probability $p$, the probability of $k$ successes is given by

$$P(k) = \binom{n}{k}p^k(1-p)^{n-k}.$$

Similarly, *multinomial coefficients* is given by

$$\binom{n}{n_1, n_2, ..., n_r} = \frac{n!}{n_1!n_2!...n_r!}.$$

## 2.6 Unordered Sampling with Replacement

Suppose that we want to sample from the set $A = \{a_1, a_2, \ldots, a_n\}$ $k$ times such that repetition is allowed and ordering does not matter. For example, if $A = \{1, 2, 3\}$ and $k = 2$, then there are 6 different ways of doing this.

How can we get the number 6 without actually listing all the possibilities? One way to think about this is to note that any of the pairs in the above list can be represented by the number of 1's, 2's and 3's it contains. That is, if $x_1$ is the number of ones, $x_2$ is the number of twos, and $x_3$ is the number of threes, we can equivalently represent each pair by a vector $(x_1, x_2, x_3)$, *i.e.,*

- $(1, 1) \to (x_1, x_2, x_3) = (2, 0, 0)$
- $(1, 2) \to (x_1, x_2, x_3) = (1, 1, 0)$
- $(2, 3) \to (x_1, x_2, x_3) = (0, 1, 1)$

Note that here $x_i \geq 0$ are integers and $x_1 + x_2 + x_3 = 2$. Thus, we can claim that the number of ways we can sample two elements from the set $A = \{1, 2, 3\}$ such that ordering does not matter and repetition is allowed is the same as solutions to the following equation

$$x_1 + x_2 + x_3 = 2,$$

where $x_i \in \{0, 1, 2\}$. We can generalize this by saying: The total number of distinct $k$ samples from an $n$-element set such that repetition is allowed and ordering does not matter is the same as the number of distinct solutions to the equation

$$x_1 + x_2 + \cdots + x_n = k,$$

where $x_i \in \{0, 1, 2, \dots\}$. The number of distinct solution to the equation is given by

$$\binom{n+k-1}{k} = \binom{n+k-1}{n-1}.$$

**Proof 1** *Let us first define following simple mapping in which we replace an integer $x_i$ with vertical lines* i.e., $|$. *For instance, $x_1 + x_2 + x_3 = 2$, then we can equivalently write $| + +|$ for $1 + 0 + 1$. We have an unique representation using vertical lines and plus signs. Each solution can be represented by $k$ vertical lines and $n - 1$ plus sings. Thus, we get*

$$\binom{n-1+k}{k} = \binom{n-1+k}{n-1}.$$

# Chapter 3

# Discrete Random Variables

## 3.1 Random Variables

A random variable $X$ is a function from the sample space to the real numbers.

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

## 3.2 Probability Mass Function (PMF)

If $X$ is a discrete random variable then its range $R_X$ is a countable set, so, we can list the elements in $R_X$. In other words, we can write

$$R_X = \{x_1, x_2, \ldots\}$$

Note that here $x_1, x_2, \ldots$ are possible values of the random variable $X$. While random variables are usually denoted by capital letters, to represent the numbers in the range we usually use lowercase letters. For a discrete random variable $X$, we are interested in knowing the probabilities of $X = x_k$.

Let $X$ be a discrete random variable with range $R_X = \{x_1, x_2, \ldots\}$ (finite or countably infinite). The function

$$P_X(x_k) = P(X = x_k), \text{ for } k = 1, 2, 3 \ldots$$

is called the *probability mass function* (PMF) of $X$. Why is it called PMF? In physics, mass is the weight over gravity:

$$m = \frac{W}{g}$$

In statistics, the probability of a discrete random variable is:

$$P(A) = \frac{n(A)}{n(all)}.$$

Thus, the weight ($W$) is analogous to the number of ways an event $A$ can occur ($n(A)$) and the gravity is analogous to the sample space ($n(all)$).

## 3.3 Special Distributions

### 3.3.1 Bernoulli Distribution

$$P_X(x) = \begin{cases} p & \text{for } x = 1 \\ 1-p & \text{for } x = 0 \\ 0 & \text{Otherwise} \end{cases}$$

A Bernoulli random variable is associated with a certain event $A$. If event $A$ occurs (for example, if you pass the test), then $X = 1$; otherwise $X = 0$. For this reason the Bernoulli random variable, is also called the *indicator* random variable.

### 3.3.2 Geometric Distribution

Suppose that I have a coin with $P(H) = p$. I toss the coin until I observe the first heads. We define $X$ as the total number of coin tosses in this experiment. Then $X$ is said to have geometric distribution with parameter $p$. In other words, you can think of this experiment as **repeating independent Bernoulli trials until observing the first success**. The range of $X$ here is $RX = \{1, 2, 3, \dots\}$.

$$P_X(k) = P(X = k) = (1-p)^{k-1}p, \text{ for } k = 1, 2, 3, \dots$$

### 3.3.3 Binomial Distribution

Suppose that I have a coin with $P(H) = p$. I toss the coin $n$ times and define $X$ to be the total number of heads that I observe. Then $X$ is binomial with parameter $n$ and $p$. The range of $X$ in this case is $R_X = 0, 1, 2, \dots, n$.

$$P_X(k) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

Here is a useful way of thinking about a binomial random variable. It can be obtained by $n$ independent coin tosses. If we think of each coin toss as a Bernoulli random variable, the $Binomial(n, p)$ random variable is a sum of $n$ independent $Bernoulli(p)$ random variables. This is stated more precisely in the following lemma.

If $X_1, X_2, \dots, X_n$ are independent $Bernoulli(p)$ random variables, then the random variable $X$ defined by $X = X_1 + X_2 + \cdots + X_n$ has a $Binomial(n, p)$ distribution.

Example:

- Let $X \sim Binomial(n, p)$ and $Y \sim Binomial(m, p)$ be two independent random variables. Define a new random variable as $Z = X + Y$. Find the PMF of $Z$.
- Solution 1: Since $X \sim Binomial(n, p)$, we can think of $X$ as the number of heads in $n$ independent coin tosses:
$$X = X_1 + \cdots + X_n,$$

where the $X_i$'s are independent Bernoulli RVs. Similarly, $Y \sim Binomial(m, p)$. Thus, the RV $Z = X + Y$ will be the total number of heads in $n + m$ coin tosses:

$$Z = X + Y = X_1 + \cdots + X_n + Y_1 + \cdots + Y_m.$$

Therefore, $Z$ is a binomial RV with parameters $m + n$ and $p$, *i.e., Binomial*$(m + n, p)$.

- Solution 2: First, we note that $R_z = \{0, 1, \ldots, m + n\}$. For $k \in R_z$, we get

$$P_Z(k) = P(Z = k) = P(X + Y = k).$$

We will find $P(X + Y = k)$ by using conditioning and the law of total probability.

$$\begin{aligned}
P(Z = k) &= P(X + Y = k) \\
&= \sum_{i=0}^{n} P(X + Y = k|X = i)P(X = i) \\
&= \sum_{i=0}^{n} P(Y = k - i|X = i)P(X = i) \\
&= \sum_{i=0}^{n} P(Y = k - i)P(X = i) \quad \text{Since } X \text{ and } Y \text{ are independent} \\
&= \sum_{i=0}^{n} \binom{m}{k-i} p^{k-i}(1-p)^{m-k+i} \binom{n}{i} p^i (1-p)^{n-i} \\
&= \sum_{i=0}^{n} \binom{m}{k-i} \binom{n}{i} p^k (1-p)^{m+n-k} \\
&= p^k (1-p)^{m+n-k} \sum_{i=0}^{n} \binom{m}{k-i} \binom{n}{i} \\
&= \binom{m+n}{k} p^k (1-p)^{m+n-k} \quad \text{by Vandermonde's identity}
\end{aligned}$$

**Negative Binomial (Pascal) Distribution** The *negative binomial* or *Pascal distribution* is a **generalization of the geometric distribution**. It relates to the random experiment of **repeated independent trials until observing $m$ successes**. Suppose that I have a coin with $P(H) = p$. I toss the coin until I observe $m$ heads, where $m \in \mathbb{N}$. We define $X$ as the total number of coin tosses in this experiment. Then $X$ is said to have Pascal distribution with parameter $m$ and $p$. We write $X \sim Pascal(m, p)$. Note that $Pascal(1, p = Geometric(p))$, since the geometric distribution repeats trials until observing the first success. Note that by our definition the range of $X$ is given by $R_X = \{m, m + 1, m + 2, m + 3, \ldots\}$, since $X$ is the number of coin tosses to observe $m$ target events.

Let's derive the PMF of a $Pascal(m, p)$ RV $X$. To find the probability of the event $A = \{X = k\}$, we argue as follows. By definition, event $A$ can be written as $A = B \cap C$, where

- $B$ is the event that we observe $m - 1$ heads (*i.e.,* successes) in the first $k - 1$ trials
- $C$ is the event that we observe a head in the $k$-th trial.

Note that $B$ and $C$ are independent events because they are related to different independent trials (coin tosses). Thus,

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

We get $P(C) = p$, so

$$P(B) = \binom{k-1}{m-1} p^{m-1}(1-p)^{(k-1)-(m-1)} = \binom{k-1}{m-1} p^{m-1}(1-p)^{k-m}.$$

Finally, we obtain

$$P(B) = \binom{k-1}{m-1} p^{m}(1-p)^{k-m}.$$

### 3.3.4 Hyper-Geometric Distribution

You have a bag that contains $b$ blue marbles and $r$ red marbles. You choose $k \leq b+r$ marbles at random (without replacement). Let $X$ be the number of blue marbles in your sample. By this definition, we have $X \leq \min(k, b)$. Also, the number of red marbles in your sample must be less than or equal to $r$, so we conclude $X \geq \max(0, k-r)$. Therefore, the range of $X$ is given by $R_X = \{\max(0, k-r), \max(0, k-r)+1, \max(0, k-r)+2, \ldots, \min(k, b)\}$.

To find $P_X(x)$, note that total number of ways to choose $k$ marbles from $b+r$ marbles is $\binom{b+r}{k}$. The total number of ways to choose $x$ blue marbles and $k-x$ red marbles is $\binom{b}{x}\binom{r}{k-x}$. Thus, we get

$$P_X(x) = \frac{\binom{b}{x}\binom{r}{k-x}}{\binom{b+r}{k}}, \qquad \text{for } x \in R_X.$$

### 3.3.5 Poisson Distribution

The Poisson distribution is one of the most widely used probability distributions. It is usually used in scenarios where we are **counting the occurrences of certain events in an interval of time or space**. In practice, it is often an approximation of a real-life random variable. Here is an example of a scenario where a Poisson random variable might be used. Suppose that we are counting the number of customers who visit a certain store from 1pm to 2pm. Based on data from previous days, we know that on average $\lambda = 15$ customers visit the store. Of course, there will be more customers some days and fewer on others. Here, we may model the random variable $X$ showing the number customers as a Poisson random variable with parameter $\lambda = 15$. Let us introduce the Poisson PMF first, and then we will talk about more examples and interpretations of this distribution.

$$P_X(k) = e^{-\lambda}\frac{\lambda^k}{k!}.$$

Note that $\lambda$ is the mean number of events within a given interval of time or space.

Example: The number of emails that I get in a weekday can be modeled by a Poisson distribution with an average of 0.2 emails per minute.

- What is the probability that I get no emails in an interval of length 5 minutes?
  - For 5 minutes, there would be 1 email on average. Thus, $\lambda = 1$,

$$P(X=0) = P_X(0) = e^{-\lambda}\frac{\lambda^0}{0!} = \frac{1}{e} \approx 0.37$$

- What is the probability that I get more than 3 emails in an interval of length 10 minutes?
  - Let $Y$ be the number of emails that I get in the 10-minute interval. Then by the assumption $Y$ is a Poisson RV with parameter $\lambda = 10 \times 0.2 = 2$. Thus,

$$\begin{aligned} P(Y > 3) &= 1 - P(Y \leq 3) \\ &= 1 - (P_Y(0) + P_Y(1) + P_Y(2) + P_Y(3)) \\ &= 1 - e^{-\lambda} - \frac{e^{-\lambda}\lambda}{1!} - \frac{e^{-\lambda}\lambda^2}{2!} - \frac{e^{-\lambda}\lambda^3}{3!} \\ &\approx 0.1429 \end{aligned}$$

Imagine you have a busy customer service center that receives phone calls. You want to know how many calls to expect in an hour, but calls can come at any moment and don't follow a strict schedule.

- Average Rate ($\lambda$): First, you determine the average number of calls you receive per hour. Let's say it's 10 calls per hour. This average rate is denoted by the symbol $\lambda$
- Probability Calculation: Using the Poisson formula, you can calculate the probability of receiving a certain number of calls in any given hour.

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

- $P(X = k)$ is the probability of getting $k$ calls in an hour.
- $e$ is the base of the natural logarithm (approximately equal to 2.71828).
- $\lambda$ is the average rate (10 calls per hour).
- $k$ is the number of calls you want to find the probability for.
- $k!$ (k factorial) is the product of all positive integers up to $k$.

### 3.3.6 Poisson as an Approximation for Binomial

The Poisson distribution can be viewed as the limit of binomial distribution. Suppose $X \sim Binomial(n, p)$ where the number of trials $n$ is very large and the probability of success $p$ is very small. In particular, assume that $\lambda = np$ is a positive constant. We show that the PMF of $X$ can be approximated by the PMF of a $Poisson(\lambda)$ random variable. The importance of this is that Poisson PMF is much easier to compute than the binomial. Let us state this as a theorem.

Let $X \sim Binomial(n, p = \frac{\lambda}{n})$, where $\lambda > 0$ is fixed. Then for any $k \in \{0, 1, 2, \dots\}$ we have

$$\lim_{n \to \infty} P_X(k) = \frac{e^{-\lambda}\lambda^k}{k!}.$$

References Poisson

## 3.4 Cumulative Distribution Function

The PMF is one way to describe the distribution of a discrete RV. As we will see later on, PMF cannot be defined for continuous random variables. The cumulative distribution function (CDF) of a random variable is another method to describe the distribution of random variables. The advantage of the CDF is that it can be defined for any kind of RV (discrete, continuous, and mixed).

**Definition 1** *Cumulative Distribution Function* The cumulative distribution function (CDF) of random variable $X$ is defined as

$$F_X(x) = P(X \leq x), \forall x \in \mathbb{R}.$$

Note that the subscript $X$ indicates that this is the CDF of the random variable $X$. Also, note that the CDF is defined for all $x \in \mathbb{R}$.

## 3.5 Expectation

If you have a collection of numbers $a_1, a_2, \ldots, a_N$, their average is a single number that describes the whole collection. Now, consider a random variable $X$. We would like to define its average, or as it is called in probability, its expected value or mean. The expected value is defined as the weighted average of the values in the range.

**Definition 2** *Expected Value* Let $X$ be a discrete RV with range $R_X = \{x_1, x_2, \ldots, \}$. The expected value of $X$, denoted by $EX$ is defined as

$$EX = \sum_{x_k \in R_X} x_k P(X = x_k) = \sum_{x \in R_X} x_k P_X(x_k)$$

## 3.6 Functions of Random Variables

If $X$ is a RV and $Y = g(X)$, then $Y$ itself is a random variable. Thus, we can talk about its PMF, CDF, and expected value. First note that the range of $Y$ can be written as

$$R_Y = \{g(x) | x \in R_X\}.$$

If we already know the PMF of $X$, to find the PMF of $Y = g(X)$, we can write

$$\begin{aligned} P_Y(y) &= P(Y = y) \\ &= P(g(X) = y) \\ &= \sum_{x : g(x) = y} P_X(x) \end{aligned}$$

**Example:** Let $X$ be a discrete RV with $P_X(k) = \frac{1}{5}$ for $k = -1, 0, 1, 2, 3$. Let $Y = 2|X|$. Find the range and PMF of $Y$. The range of $Y$ is

$$\begin{aligned} R_Y &= \{2|x|\} \\ &= \{0, 2, 4, 6\} \end{aligned}$$

To find $P_Y(y)$, we need to find $P(Y = y)$ for $y = 0, 2, 4, 6$:

$$\begin{aligned} P_Y(0) &= P(Y = 0) = P(2|x| = 0) \\ &= P(X = 0) = \frac{1}{5} \\ P_Y(2) &= P(Y = 2) = P(2|x| = 2) \\ &= P(X = -1 \text{ or } X = 1) \\ &= P_X(-1) + P_X(1) = \frac{2}{5} \\ &\vdots \end{aligned}$$

### 3.6.1 Expected Value of a Function of a Random Variable (LOTUS)

Let $X$ be a discrete random variable with PMF $P_X(x)$, and let $Y = g(X)$. Suppose that we are interested in finding $EY$. One way to find $EY$ is to first find the PMF of $Y$ and then use the expectation formula $EY = E[g(X)] = \sum_{y \in R_y} y P_Y(y)$. But there is another way which is usually easier. It is called the *law of the unconscious statistician* (LOTUS).

$$\mathbb{E}[g(X)] = \sum_{x_k \in R_X} g(x_k) P_X(x_k)$$

One of the main points of the theorem is that you can compute $\mathbb{E}[g(X)]$ without computing $P_Y(y)$. In practice it is usually easier to use LOTUS than direct definition when we need $\mathbb{E}[g(X)]$.

## 3.7 Variance

The variance of a random variable $X$, with mean $EX = \mu_X$, is defined as

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

To compute $\text{Var}(X) = \mathbb{E}[(X - \mu_X)^2]$, note that we need to find the expected value of $g(X) = (X - \mu_X)^2$, so we can use **LOTUS**. In particular, we can write

$$\text{Var}(X) = \mathbb{E}[(X - \mu_X)^2] = \sum_{x_k \in R_X} (x_k - \mu_X)^2 P_X(x_k).$$

## 3.8 Standard Deviation

$$\text{SD}(X) = \sigma_X = \sqrt{\text{Var}(X)}.$$

A useful formula for computing the variance is

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

We can find $\mathbb{E}[X^2]$ using LOTUS:

$$\mathbb{E}[X^2] = \sum_{x_k \in R_X} x_k^2 P_X(x_k).$$

# Chapter 4

# Continuous and Mixed Random Variables

## 4.1 Introduction

Remember that discrete random variables can take only a countable number of possible values. On the other hand, a continuous random variable $X$ has a range in the form of an interval or a union of non-overlapping intervals on the real line (possibly the whole real line). Also, for any $x \in \mathbb{R}$, $P(X = x) = 0$. Thus, we need to develop new tools to deal with continuous random variables. The good news is that the theory of continuous random variables is completely analogous to the theory of discrete random variables. Indeed, if we want to oversimplify things, we might say the following: take any formula about discrete random variables, and then replace sums with integrals, and replace PMFs with probability density functions (PDFs), and you will get the corresponding formula for continuous random variables.

## 4.2 Probability Density Function (PDF)

To determine the distribution of a discrete random variable we can either provide its PMF or CDF. For continuous random variables, the CDF is well-defined so we can provide the CDF. However, **the PMF does not work for continuous random variables, because for a continuous random variable** $P(X = x) = 0$, $\forall x \in \mathbb{R}$. Instead, we can usually define the *probability density function* (PDF). The PDF is the *density* of probability rather than the probability mass. The concept is very similar to mass density in physics: its unit is probability per unit length. For example, let the bus waiting time be uniformly distributed: $X \sim [10, 30]$. The probability of waiting between 15 and 20 minutes is:

$$P(X < 15) = \int_{15}^{20} \frac{1}{20}\, dx = \frac{1}{20} \cdot (20 - 15) = \frac{1}{4} = 0.25.$$

So, the mass is analogous to the interval ($[a, b] = [15, 20]$) and the volume is analogous to the entire range ($[c, d] = [10, 30]$). To get a feeling for PDF, consider a continuous random variable $X$ and define the function $f_X(x)$ as follows (wherever the limit exists):

$$f_X(x) = \lim_{\Delta \to 0^+} \frac{P(x < X \leq x + \Delta)}{\Delta}.$$

The function $f_X(x)$ gives us the probability density at point $x$. It is the limit of the probability of the interval $(x, x+\Delta]$ divided by the length of the interval as the length of the interval goes to 0. Remember that

$$P(x < X \leq x+\Delta) = F_X(x+\Delta) - F_X(x).$$

Thus, we get

$$\begin{aligned}
f_X(x) &= \lim_{\Delta \to 0} \frac{F_X(x+\Delta) - F_X(x)}{\Delta} \\
&= \frac{dF_X(x)}{dx} \\
&= F'_X(x), \quad \text{if } F_X(x) \text{ is differentiable at } x.
\end{aligned}$$

Let's find the PDF of the uniform random variable $X \sim Uniform(a,b)$, which can be expressed as follows:

$$f_X(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & x < a \text{ or } x > b \end{cases}$$

Note that the CDF is not differentiable at points $a$ and $b$. Nevertheless, this is not important at this moment.

The uniform distribution is the simplest continuous random variable you can imagine. For other types of continuous random variables the PDF is non-uniform. Note that for small values of $\delta$ we can write

$$P(x < X \leq x+\delta) \approx f_X(x)\delta.$$

Thus, if $f_X(x_1) < f_X(x_2)$, we can say $P(x_1 < X \leq x_1+\delta) < P(x_2 < X \leq x_2+\delta)$ , *i.e.,* the value of $X$ is more likely to be around $x_2$ than $x_1$.

Since the PDF is the derivative of the CDF, the CDF can be obtained from PDF by integrations (by assuming absolute continuity):

$$F_X(x) = \int_{-\infty}^{x} f_X(u)du.$$

Also, we have

$$P(a < X \leq b) = F_X(b) - F_X(a) = \int_a^b f_X(u)du.$$

More generally, for a set $A$, $P(X \in A) = \int_a^b f_X(u)du$. Note that if we integrate over the entire real line, we must get 1, *i.e.,*

$$\int_{-\infty}^{\infty} f_X(u)du = 1.$$

## 4.3 Expected Value and Variance

As we mentioned earlier, the theory of continuous random variables is very similar to the theory of discrete random variables. In particular, usually summations are replaced by integrals and PMFs are replaced by PDFs. The proofs and ideas are very analogous to the discrete case, so sometimes we state the results without mathematical derivations for the purpose of brevity.

Recall that the expected value of a discrete random variable can be obtained as

$$EX = \sum_{x_k \in R_X} x_k P_X(x_k).$$

The expected value of a continuous RV as

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

### 4.3.1 Expected Value of a Function of a Continuous Random Variable

Law of the unconscious statistician (LOTUS) for continuous random variables:

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

### 4.3.2 Variance

$$\begin{aligned}
\mathrm{Var}(X) &= \mathbb{E}[(X - \mu_X)^2] \\
&= \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x) dx \\
&= EX^2 - (EX)^2 \\
&= \int_{-\infty}^{\infty} x^2 f_X(x) dx - \mu_X^2
\end{aligned}$$

Note that for $a, b \in \mathbb{R}$, we always have

$$\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X)$$

## 4.4 Functions of Continuous Random Variables

If $X$ is a continuous random variable and $Y = g(X)$ is a function of $X$, then $Y$ itself is a random variable. Thus, we should be able to find the CDF and PDF of $Y$. It is usually more straightforward to start from the CDF and then to find the PDF by taking the derivative of the CDF. Note that before differentiating the CDF, we should check that the CDF is continuous. As we will see later, the function of a continuous random variable might be a non-continuous random variable. Let's look at an example.

**Example:** Let $X$ be a $Uniform(0, 1)$ random variable, and let $Y = e^X$.

- CDF of $Y$
- PDF of $Y$
- $EY$

The PDF of $X$ is given by

$$F_X(x) = \begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } 0 \leq x \leq 1 \\ 1 & \text{for } x > 1 \end{cases}$$

The range of $x$, $R_X = [0, 1]$, so the range of $Y$, $R_Y = [1, e]$. We can find the CDF of $Y$ as follows:

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P(e^X \leq y) \\ &= P(X \leq \ln y) \quad \text{, since } e^x \text{ is an increasing function.} \\ &= F_X(\ln y) \quad \text{by definition.} \\ &= \ln y \quad \text{, since } F_X(x) = x \text{ for } 0 \leq x \leq 1 \text{ and } 0 \leq \ln y \leq 1. \end{aligned}$$

## 4.5 The Method of Transformations

So far, we have discussed how we can find the distribution of a function of a continuous random variable starting from finding the CDF. If we are interested in finding the PDF of $Y = g(X)$, and the function $g$ satisfies following properties, it might be easier to use a method called the method of transformations.

- $g(x)$ is differentiable;
- $g(x)$ is a strictly increasing function, that is, if $x_1 < x_2$, then $g(x_1) < g(x_2)$.

Now, let $X$ be a continuous random variable and $Y = g(X)$. We will show that you can directly find the PDF of $Y$ using the following formula.

$$f_Y(y) = \begin{cases} \frac{f_X(x_1)}{g'(x_1)} = f_X(x_1).\frac{dx_1}{dy} & \text{where } g(x_1) = y \\ 0 & \text{if } g(x) = y \text{ does not have a solution} \end{cases}$$

Note that the derivative $\frac{dx}{dy}$ or $\frac{d}{dy}(g^{-1}(y))$ **measures how $X$ changes with respect to $Y$**. Since $g$ is strictly increasing, its inverse function $g^{-1}$ is well defined. That is, for each $y \in R_Y$, there exists a **unique** $x_1$ such that $g(x_1) = y$. We can write $x_1 = g^{-1}(y)$.

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P(g(X) \leq y) \\ &= P(X < g^{-1}(y)) \quad \text{since } g \text{ is strictly increasing.} \\ &= F_X(g^{-1}(y)). \end{aligned}$$

To find the PDF of $Y$, we differentiate $F_Y(y)$ as follows:

$$\begin{aligned}
f_Y(y) &= \frac{d}{dy} F_X(x_1) \quad \text{by } g(x_1) = y \\
&= \frac{dx_1}{dy} \cdot \underbrace{\frac{d}{dx_1} F_X(x_1)}_{=F'_X(x_1)} \\
&= \frac{dx_1}{dy} f_X(x_1) \\
&= f_X(g^{-1}(y)) \left| \frac{d}{dy}(g^{-1}(y)) \right|
\end{aligned}$$

We can repeat the same argument for the case where $g$ is **strictly decreasing**. In that case, $g'(x_1)$ will be **negative**, so we need to use $|g(x_1)|$ . Thus, we can state the following theorem for a *strictly monotonic function*. (A function $g : R \to R$ is called strictly monotonic if it is strictly increasing or strictly decreasing.)

Actually, we assumed that $g$ was one-to-one out of convenience: the condition that $g$ is one-to-one is not necessary for change of variables to work: Consider a continuous random variable $X$ with domain $R_X$, and let $Y = g(X)$. Suppose that we can partition $R_X$ into a finite number of intervals such that $g(x)$ is strictly monotone and differentiable on each partition. Then the PDF of $Y$ is given by

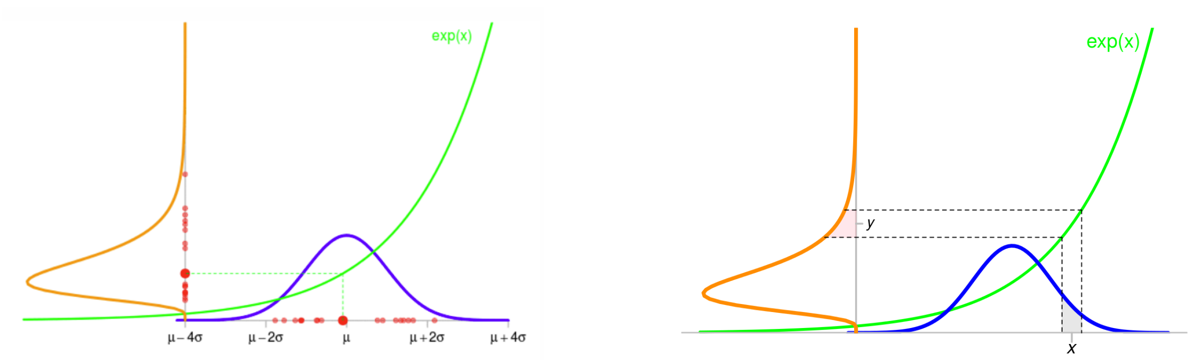
$$f_Y(y) = \sum_{i=1}^{n} \frac{f_X(x_i)}{|g'(x_i)|} = \sum_{i=1}^{n} f_X(x_i). \left| \frac{dx_i}{dy} \right|,$$

where $x_1, \ldots, x_n$ are real solutions to $g(x) = y$.

### 4.5.1 Intuitive Explanation

"How to derive the PDF of the random variable $Y = g(X)$ when one knows the pdf of the random variable $X$?". For a general function $g$, there is no direct formula to get the pdf of the random variable $Y = g(X)$ knowing the PDF of $X$. There is a formula in case when $h$ is a differentiable one-to-one mapping from the range (the support, I should say) of $X$ to the range of $Y$.

Take for example a random variable $X \sim \mathcal{N}(\mu, \sigma)$ and set $Y = \exp(X)$. The figure below shows some simulations of $X$ and the corresponding values of $Y$. The density of $X$ is shown in blue and the one of $Y$ is shown in orange in the vertical direction. Now the question is: knowing the




density of $X$, what is the density of $Y$? Taking a point $y$ in the range of $Y$, the PDF $f_Y$ provides the probability of $Y$, belong to a small area $dy$ around $y$ by the formula below

$$P(Y \in dy) \approx f_Y(y)|dy|,$$

where $P(Y \in dy)$ is the area below the curve. Similarly, we can define

$$P(X \in dx) \approx f_X(x)|dx|$$

The above two areas are approximately the same in case of very small region. Note that if $dy$ and $dx$ are very small, we can approximate the derivative of $g'(x) = \frac{|dy|}{|dx|}$. Compactly, this can be expressed as follows:

$$P(Y \in dy) = P(X \in dx) = f_X(x)\frac{|dy|}{g'(x)}$$

With $y = g(x)$ we can get

$$\begin{aligned}
P(Y \in dy) &= f_X(x)\frac{|dy|}{g'(x)} \\
&= f_X(g^{-1}(y))\frac{|dy|}{g'(g^{-1}(y))} \\
&= f_X(g^{-1}(y))|dy|(g^{-1})'(y)
\end{aligned}$$

The last line is by the derivative of inverse function which is

$$\frac{d}{dx}f^{-1}(x) = \frac{1}{f'(f^{-1}(x))}$$

Finally, we can get

$$f_Y(y) = f_X(g^{-1}(y))|(g^{-1})'(y)|$$

Note that the absolute is determined by the function $h$. This is the so-called change of variables formula.

## 4.6 Various Distributions

### 4.6.1 Uniform Distribution

$$f_X(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & x < a \text{ or } x > b \end{cases}$$

### 4.6.2 Exponential Distribution

Check more in Exponential Dist.

A continuous random variable $X$ is said to have an exponential distribution with parameter $\lambda > 0$, shown as $X \sim \mathit{Exponential}(\lambda)$, if its PDF is given by

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

### 4.6.3 Gamma Distribution

The gamma distribution is another widely used distribution. Its importance is largely due to its relation to exponential and normal distributions. Here, we will provide an introduction to the gamma distribution. Before introducing the gamma random variable, we need to introduce the gamma function.

**Gamma function** $\Gamma(x)$ is an extension of the factorial function to real (and complex) numbers. In specific, if $n \in \{1, 2, 3, \dots\}$, then

$$\Gamma(n) = (n-1)!.$$

More generally, for any positive real number $\alpha$, $\Gamma(\alpha)$ is defined as follows:

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx, \text{ for } \alpha > 0.$$

Note that for $\alpha = 1$,

$$\Gamma(\alpha) = 1.$$

**Gamma Distribution** is a distribution with parameters $\alpha > 0$ and $\lambda > 0$. its PDF is given by

$$f_X(x) = \begin{cases} \frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

## 4.7 Gaussian Distribution

The normal distribution is by far **the most important probability distribution**. One of the main reasons for that is the *Central Limit Theorem* (CLT). To give you an idea, the CLT states that *if you add a large number of random variables, the distribution of the sum will be approximately normal under certain conditions.* The importance of this result comes from the fact that many random variables in real life can be expressed as the sum of a large number of random variables and, by the CLT, we can argue that distribution of the sum should be normal. The CLT is one of the most important results in probability and we will discuss it later on. Here, we will introduce normal random variables.

We first define the standard normal random variable. We will then see that we can obtain other normal random variables by scaling and shifting a standard normal random variable.

A continuous random variable $Z$ is said to be a *standard normal* (*standard Gaussian*) random variable, shown as $Z \sim \mathcal{N}(0, 1)$, if its PDF is given by

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{z^2}{2}\right\}, \quad \text{for all } z \in \mathbb{R}.$$

The $1/\sqrt{2\pi}$ is there to make sure that the area under the PDF is equal to one.

### 4.7.1 Cumulative Distribution Function

The CDF of the standard normal distribution is denoted by the $\Phi$ function:

$$\Phi(x) = P(Z \leq x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left\{-\frac{u^2}{2}\right\} du.$$

Here are some properties of the $\Phi$ function:

- $\lim_{x\to\infty} \Phi(x) = 1$
- $\lim_{x\to-\infty} \Phi(x) = 0$
- $\Phi(0) = \frac{1}{2}$
- $\Phi(-x) = 1 - \Phi(x), \forall x \in \mathbb{R}.$

### 4.7.2 Multinomial

For a $D$-dimensional vector $\mathbf{x}$, the multivariate Gaussian distribution takes the form

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left( -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \tag{4.1}$$

$$\tag{4.2}$$

### 4.7.3 Conditional Gaussian Distribution

Consider first the case of conditional distributions. Suppose $\mathbf{x}$ is a $D$-dimensional vector with Gaussian distribution $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and that we partition $\mathbf{x}$ into two disjoint subsets $\mathbf{x}_a$ and $\mathbf{x}_b$. Thus, $\mathbf{x}_a$ has $M$ components and $\mathbf{x}_b$ has $D - M$ components.

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{bmatrix}.$$

Similarly,

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{bmatrix}$$

and the covariance matrix is given by

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{bmatrix}$$

Note that the symmetry $\boldsymbol{\Sigma}^T = \boldsymbol{\Sigma}$ implies that $\boldsymbol{\Sigma}_{ab}^T = \boldsymbol{\Sigma}_{ba}$. We can also define a *precision matrix* as follows:

$$\boldsymbol{\Lambda} \equiv \boldsymbol{\Sigma}^{-1}$$

We also introduce a partitioned form of the precision matrix:

$$\boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{bmatrix} \tag{4.3}$$

Because the inverse of a symmetric matrix is also symmetric, we see that $\mathbf{\Lambda}_{aa}$ and $\mathbf{\Lambda}_{bb}$ are symmetric and $\mathbf{\Lambda}_{ab}^T = \mathbf{\Lambda}_{ba}$. Note that, for instance, $\mathbf{\Lambda}_{aa}$ is not simply given by the inverse of $\mathbf{\Sigma}_{aa}$.

Now let's compute the conditional probability:

$$\begin{aligned}
-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) &= \frac{1}{2}\left(\left(\begin{bmatrix}\mathbf{x}_a\\ \mathbf{x}_b\end{bmatrix} - \begin{bmatrix}\boldsymbol{\mu}_a\\ \boldsymbol{\mu}_b\end{bmatrix}\right)^T \begin{bmatrix}\mathbf{\Lambda}_{aa} & \mathbf{\Lambda}_{ab}\\ \mathbf{\Lambda}_{ba} & \mathbf{\Lambda}_{bb}\end{bmatrix}\left(\begin{bmatrix}\mathbf{x}_a\\ \mathbf{x}_b\end{bmatrix} - \begin{bmatrix}\boldsymbol{\mu}_a\\ \boldsymbol{\mu}_b\end{bmatrix}\right)\right) \\
&= -\frac{1}{2}\big((\mathbf{x}_a-\boldsymbol{\mu}_a)^T\mathbf{\Lambda}_{aa}(\mathbf{x}_a-\boldsymbol{\mu}_a) + (\mathbf{x}_a-\boldsymbol{\mu}_a)^T\mathbf{\Lambda}_{ab}(\mathbf{x}_a-\boldsymbol{\mu}_a) \\
&\quad + (\mathbf{x}_a-\boldsymbol{\mu}_a)^T\mathbf{\Lambda}_{ba}(\mathbf{x}_a-\boldsymbol{\mu}_a) + (\mathbf{x}_a-\boldsymbol{\mu}_a)^T\mathbf{\Lambda}_{bb}(\mathbf{x}_a-\boldsymbol{\mu}_a)\big)
\end{aligned}$$

## 4.8 Mixed Random Variables

The mixed random variables are random variables that are **neither discrete nor continuous, but are a mixture of both**.

To find the cumulative distribution function (CDF) of $Y$, given that $Y = g(X)$ and the transformation $g(X)$ is defined as:

$$g(X) = \begin{cases} X^2 & 0 \le X \le \frac{1}{2} \\ 2X - 1 & \frac{1}{2} < X \le 1 \end{cases}$$

1. Determine the PDF of $X$. The given PDF of $X$ is:

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

2. Determine the ranges of $Y$. The ranges of $Y$ are derived from the transformation:

   - For $0 \le X \le \frac{1}{2}$:

$$Y = X^2$$
$$0 \le Y \le \left(\frac{1}{2}\right)^2 = \frac{1}{4}$$

   - For $\frac{1}{2} < X \le 1$:

$$Y = 2X - 1$$
$$2\left(\frac{1}{2}\right) - 1 < Y \le 2(1) - 1$$
$$0 < Y \le 1$$

3. Combining these, we get the range of $Y$ as $0 \le Y \le 1$. Find the CDF of $Y$. The CDF of $Y$, $F_Y(y)$, is given by $F_Y(y) = P(Y \le y)$. We need to consider the two different transformations:

   (a) For $0 \le y \le \frac{1}{4}$:

$$Y = X^2$$
$$P(Y \le y) = P(X^2 \le y) = P(X \le \sqrt{y})$$

$$F_Y(y) = P(X \leq \sqrt{y}) = \int_0^{\sqrt{y}} 2x\,dx$$

$$F_Y(y) = \left[x^2\right]_0^{\sqrt{y}} = (\sqrt{y})^2 = y$$

(b) For $\frac{1}{4} < y \leq 1$:

$$Y = 2X - 1$$

$$P(Y \leq y) = P(2X - 1 \leq y) = P(X \leq \frac{y+1}{2})$$

$$F_Y(y) = P(X \leq \frac{y+1}{2}) = \int_0^{\frac{y+1}{2}} 2x\,dx$$

$$F_Y(y) = \left[x^2\right]_0^{\frac{y+1}{2}} = \left(\frac{y+1}{2}\right)^2$$

$$F_Y(y) = \frac{(y+1)^2}{4}$$

4. Combining these results, the CDF of $Y$ is:

$$F_Y(y) = \begin{cases} y & 0 \leq y \leq \frac{1}{4} \\ \frac{(y+1)^2}{4} & \frac{1}{4} < y \leq 1 \end{cases}$$

### 4.8.1 Delta Function

In this section, we will **use the Dirac delta function to analyze mixed random variables**. Technically speaking, *the Dirac delta function is not actually a function.* It is what we may call a generalized function. Nevertheless, its definition is intuitive and it simplifies dealing with probability distributions.

Remember that any random variable has a CDF. Thus, we can use the CDF to answer questions regarding discrete, continuous, and mixed random variables. On the other hand, the PDF is defined only for continuous random variables, while the PMF is defined only for discrete random variables. Using **delta functions will allow us to define the PDF for discrete and mixed random variables.** Thus, it allows us to unify the theory of discrete, continuous, and mixed random variables.

**Dirac Delta Function** We cannot define the PDF for a discrete random variable because its CDF has jumps. If we could somehow differentiate the CDF at jump points, we would be able to define the PDF for discrete random variables as well. This is the idea behind our effort in this section. Here, we will introduce the *Dirac delta function* and discuss its application to probability distributions. Let's derive the Dirac delta function.

First, consider the following unit step function $u(x)$:

$$u(x) = \begin{cases} 1 & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

This function has a discontinuity at $x = 0$. Let us remove the jump and define, for any $\alpha > 0$, the function $u_\alpha(x)$ as

$$u_\alpha(x) = \begin{cases} 1 & x > \frac{\alpha}{2} \\ \frac{1}{\alpha}(x + \frac{\alpha}{2}) & -\frac{\alpha}{2} \leq x \leq \frac{\alpha}{2} \\ 0 & x < -\frac{\alpha}{2} \end{cases}$$

The good thing about $u_\alpha(x)$ is that it is a continuous function. Now let us define the function $\delta_\alpha(x)$ as the derivative of $u_\alpha(x)$ wherever it exists.

$$\delta_\alpha(x) = \frac{du_\alpha(x)}{dx} = \begin{cases} \frac{1}{\alpha} & |x| < \frac{\alpha}{2} \\ 0 & |x| > \frac{\alpha}{2} \end{cases}$$

We can notice that

$$\delta_\alpha(x) = \frac{d}{dx}u_\alpha(x), \quad u(x) \overset{\text{a.e.}}{=} \lim_{\alpha \to 0} u_\alpha(x)^1$$

Now, we would like to define the delta "function", $\delta(x)$, as

$$\delta(x) = \lim_{\alpha \to 0} \delta_\alpha(x).$$

Note that as $\alpha$ becomes smaller and smaller, the height of $\delta_\alpha(x)$ becomes larger and larger and its width becomes smaller and smaller. Taking the limit, we obtain

$$\delta(x) = \begin{cases} \infty & x = 0 \\ 0 & \text{otherwise} \end{cases}$$

Equivalently,

$$\delta(x) = \frac{d}{dx}u(x).$$

Intuitively, with extremely small $\alpha$, we would like to have the following definitions. Let $g : \mathbb{R} \to \mathbb{R}$ be a continuous function. We define

$$\int_{-\infty}^{\infty} g(x)\delta(x - x_0)dx = \lim_{\alpha \to 0} \left[ \int_{-\infty}^{\infty} g(x)\delta_\alpha(x - x_0)dx \right]$$

Then, we have the following lemma, which in fact is the most useful property of the delta function.

Let $g : \mathbb{R} \to \mathbb{R}$ be a continuous function. We have

$$\int_{-\infty}^{\infty} g(x)\delta(x - x_0)dx = g(x_0).$$

**Using the Delta Function in PDFs of Discrete and Mixed RV** Consider a discrete random variable $X$ with range $R_X = \{x_1, \ldots,_n\}$ and PMF $P_X(x_k)$. Note that the CDF for X can be written as

$$F_X(x) = \sum_{x_k \in R_X} P_X(x_k)u(x - x_k).$$

where:

[1]The term almost everywhere is abbreviated a.e.; in older literature p.p. is used, to stand for the equivalent French language phrase presque partout.

- $u(x - x_k)$ is the Heaviside step function, which is defined as:

$$u(x - x_k) = \begin{cases} 0 & \text{if } x < x_k \\ 1 & \text{if } x \geq x_k \end{cases}$$

The sum $\sum_{x_k \in \mathbb{R}_X} P_X(x_k)u(x - x_k)$ effectively includes only those $x_k$ values that are less than or equal to $x$ due to the step function $u(x - x_k)$. Therefore, it accumulates the probabilities $P_X(x_k)$ for all $x_k \leq x$.

Now that we have symbolically defined the derivative of the step function as the delta function, we can write a PDF for $X$ by "differentiating" the CDF:

$$\begin{aligned} f_X(x) &= \frac{dF_X(x)}{dx} \\ &= \sum_{x_k \in R_X} P_X(x_k) \frac{d}{dx} u(x - x_k) \\ &= \sum_{x_k \in R_X} P_X(x_k) \delta(x - x_k) \end{aligned}$$

We call this the **generalized PDF**.

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

$$\begin{aligned} EX &= \int_{-\infty}^{\infty} x f_X(x) dx \\ &= \int_{-\infty}^{\infty} x \sum_{x_k \in R_X} P_X(x_k) \delta(x - x_k) dx \\ &= \sum_{x_k \in R_X} P_X(x_k) \int_{-\infty}^{\infty} x \delta(x - x_k) dx \\ &= \sum_{x_k \in R_X} x_k P_X(x_k) \end{aligned}$$

# Chapter 5

# Joint Distributions

In real life, we are often **interested in several random variables that are related to each other.** For example, suppose that we choose a random family, and we would like to study the number of people in the family, the household income, the ages of the family members, etc. Each of these is a random variable, and we suspect that they are dependent. In this chapter, we develop tools to study joint distributions of random variables. The concepts are similar to what we have seen so far. The only difference is that instead of one random variable, we consider two or more. In this chapter, we will focus on two random variables, but once you understand the theory for two random variables, the extension to $n$ random variables is straightforward. We will first discuss joint distributions of discrete random variables and then extend the results to continuous random variables.

## 5.1 Joint PMF

Recall that for a discrete RV $X$, we define the PMF as $P_X(x) = P(X = x)$. Now, if we have two RVs $X$ and $Y$, we define the joint PMF as follows:

$$P_{XY}(x, y) = P(X = x, Y = y).$$

Note that the comma means "and", so we can write as

$$\begin{aligned} P_{XY}(x, y) &= P(X = x, Y = y) \\ &= P(X = x \text{ and } Y = y) \\ &= P(X = x \cap Y = y) \end{aligned}$$

We can define the joint range for $X$ and $Y$ as

$$R_{XY} = \{(x, y) | P_{XY}(x, y) > 0\}.$$

In particular, if $R_X = \{x_1, x_2, \dots\}$ and $R_Y = \{y_1, y_2, \dots\}$, then

$$\begin{aligned} R_{XY} &\subset R_X \times R_Y \\ &= R_{XY} = \{(x_i, y_j) | x_i \in R_X, y_j \in R_Y\}. \end{aligned}$$

For two discrete RVs, we have

$$\sum_{(x_i, y_j) \in R_{XY}} P_{XY}(x_i, y_j) = 1$$

We can use the joint PMF to find $P((X,Y) \in A)$ for any set $A \subset \mathbb{R}^2$. Specifically, we have

$$P((X,Y) \in A) = \sum_{(x_i,y_j)\in(A\cap R_{XY})} P_{XY}(x_i,y_j)$$

## 5.2 Joint CDF

The joint cumulative distribution function of two random variables $X$ and $Y$ is defined as

$$F_{XY}(x,y) = P(X \le x, Y \le y).$$

Equivalently,

$$\begin{aligned} F_{XY}(x,y) &= P(X \le x, Y \le y) \\ &= P(X \le x \cap Y \le y) \end{aligned}$$

If we know the CDF of $X$ and $Y$, we can find the *marginal* CDFs, $F_X(x)$ and $F_Y(y)$. Specifically, for any $x \in \mathbb{R}$, we have

$$\begin{aligned} F_{XY}(x,\infty) &= P(X \le x, Y \le \infty) \\ &= P(X \le x) \\ &= F_X(x) \end{aligned}$$

## 5.3 Conditioning and Independence

$$P(A|B) = \frac{P(A\cap B)}{P(B)}, \text{ when } P(B) > 0.$$

### 5.3.1 Conditional PMF and CDF

The conditional PMF of $X$ given an event $A$ is given by

$$\begin{aligned} P_{X|A}(x_i) &= P(X = x_i|A) \\ &= \frac{P(X = x_i \text{ and } A)}{P(A)} \end{aligned}$$

Similarly,

$$F_{X|A}(x) = P(X \le x|A)$$

### 5.3.2 Conditional PMF of $X$ given $Y$

We have observed the value of a random variable $Y$, and we need to update the PMF of another random variable $X$ whose value has not yet been observed. In these problems, we use the conditional PMF of $X$ given $Y$:

$$\begin{aligned} P_{X|Y}(x_i|y_j) &= P(X = x_i|Y = y_j) \\ &= \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)} \\ &= \frac{P_{XY}(x_i, y_j)}{P_Y(y_j)} \end{aligned}$$

### 5.3.3 Independent Random Variables

Two discrete RVs $X$ and $Y$ are independent if

$$P_{XY}(x, y) = P_X(x)P_Y(y), \forall x, y.$$

Equivalently,

$$F_{XY}(x, y) = F_X(x)F_Y(y), \forall x, y.$$

If $X$ and $Y$ are independent,

$$P_{X|Y}(x_i|y_j) = P_X(x_i).$$

### 5.3.4 Conditional Expectation

Given that we know an event $A$ has occurred, we can compute the conditional expectation of a RV $X$, $E[X|A]$:

$$E[X|A] = \sum_{x_i \in R_X} x_i P_{X|A}(x_i).$$

Similarly, given that we have observed the value of random variable $Y$, we can compute the conditional expectation of $X$:

$$E[X|Y = y] = \sum_{x_i \in R_X} x_i P_{X|Y}(x_i|y).$$

## 5.4 The Law of Total Probability

Recall that the law of total probability: If $B_1, B_2, \dots$ is a partition of the sample space $S$, then for any event $A$ we have

$$P(A) = \sum_i P(A \cap B_i) = \sum_i P(A|B_i)P(B_i).$$

If $Y$ is a discrete random variable with range $R_y = \{y_1, y_2, \dots\}$, then the events $\{Y = y_1\}, \{Y = y_2\}, \dots,$ form a partition of the sample space. Thus, we can use the law of total probability:

$$P_X(x) = \sum_{y_j \in R_Y} P_{XY}(x, y_j) = \sum_{y_j \in R_Y} P_{X|Y}(x|y_j)P_Y(y_j).$$

We can write this more generally as

$$P(X \in A) = \sum_{y_j \in R_Y} P(X \in A|Y = y_j)P_Y(y_j), \text{ for any set} A.$$

Similarly, we can write the law of total expectation:

$$EX = \sum_i E[X|B_i]P(B_i)$$
$$EX = \sum_{y_j \in R_Y} E[X|Y = y_i]P_Y(y_j).$$

This means that the expected value of $X$ can be calculated from the probability distribution of $X|Y$ and $Y$, which is often useful both in theory and practice.

## 5.5 Functions of Two Random Variables

Suppose that you have two discrete random variables $X$ and $Y$, and suppose that $Z = g(X, Y)$, where $g : \mathbb{R}^2 \to \mathbb{R}$. Then, the PMF of $Z$ is given by

$$\begin{aligned} P_Z(z) &= P(g(X, Y) = z) \\ &= \sum_{(x_i, y_j) \in A_z} P_{XY}(x_i, y_j), \end{aligned}$$

where $A_z = \{(x_i, y_j) \in R_{XY} : g(x_i, y_j) = z\}$. Note that if we are only interested in $E[g(X, Y)]$, we can directly use LOTUS, without finding $P_Z(z)$:

$$E[g(X, Y)] = \sum_{(x_i, y_j) \in R_{XY}} g(x_i, y_j)P_{XY}(x_i, y_j).$$

## 5.6 Conditional Expectation and Conditional Variance

### 5.6.1 Conditional Expectation as a Function of a Random Variable

Note that

- $E[X]$ is a scalar value
- $E[X|Y]$ is a random variable, because the value depends on $Y$.

$$E[X] = \sum_x x \cdot p(x)$$
$$E[E[X|Y]] = E[X]$$

Since, $E[E[X|Y]]$ is the function of $Y$. It is also called *the law of iterated expectations.*

$$\begin{aligned}
E[E[X|Y]] &= E\left[\sum_x x \cdot P(X = x|Y)\right] \\
&= \sum_y \left[\sum_x x \cdot P(X = x|Y)\right] \cdot P(Y = y) \\
&= \sum_y \sum_x x \cdot P(X = x, Y) \\
&= \sum_x x \sum_y \cdot P(X = x, Y) \\
&= \sum_x x \cdot P(X = x) \\
&= E[X]
\end{aligned}$$

$$\begin{aligned}
E[Y \mid X = x] &= \sum_y y \cdot p_{Y|X}(y \mid X = x) \\
&= \sum_y y \cdot \frac{p_{X,Y}(x, y)}{p_X(x)} \\
&= \sum_y y \cdot \frac{\sum_z p_{X,Y,Z}(x, y, z)}{p_X(x)} \\
&= \sum_y y \cdot \frac{\sum_z p_{Y|X,Z}(y \mid X = x, Z = z) \cdot p_{X,Z}(x, z)}{p_X(x)} \\
&= \sum_z \frac{p_{X,Z}(x, z)}{p_X(x)} \sum_y y \cdot p_{Y|X,Z}(y \mid X = x, Z = z) \\
&= \sum_z p_{Z|X}(z \mid X = x) \cdot \sum_y y \cdot p_{Y|X,Z}(y \mid X = x, Z = z) \\
&= \sum_z p_{Z|X}(z \mid X = x) \cdot E[Y \mid X = x, Z = z] \\
&= E\left[E[Y \mid X, Z] \mid X = x\right]
\end{aligned}$$

Note that if $X$ and $Y$ are independent,

- $E[X|Y] = EX$.
- $E[g(X)|Y] = E[g(X)]$.
- $E[XY] = EXEY$.
- $E[g(X)h(Y)] = E[g(X)]E[h(Y)]$.

### 5.6.2 Conditional Variance

We can define the conditional variance of $X$, $Var(X|Y = y)$:

# Bibliography

[1] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[2] Rachit Singh. . `https://rachitsingh.com/elbo_surgery/`, 2017. Online; accessed 29 January 2014.

[3] Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. Achieving human parity on automatic chinese to english news translation. *CoRR*, abs/1803.05567, 2018.