Random Signal Processing (ET2202)

Introduction

Course Structure

- Introduction
- Random Variables
- Bivariate Random Variables
- Random Vectors
- Random Vectors
- Random Processes

Introduction to Signals

- Introduction
- Random Variables
- Bivariate Random Variables
- Random Vectors
- Random Vectors
- Random Processes

Signals

- Signals are variables that carry information.
- It is described as a function of one or more independent variables.
- Basically it is a physical quantity.
- It varies with some independent or dependent variables.
- Signals can be One-dimensional or multidimensional.
- **Signal**: A function of one or more variables that convey information on the nature of a physical phenomenon.
  - Examples: v(t),i(t),x(t),heartbeat, blood pressure, temperature, vibration.

- **One-dimensional signals**: function depends on a single variable, e.g., speech signal

- **Multi-dimensional signals**: function depends on two or more variables, e.g., image

Classification of Signals

1. Continuous-time and discrete-time signals
2. Periodic and non-periodic signals
3. Casual and Non-casual signals
4. Deterministic and random signals

Continuous vs. Discrete

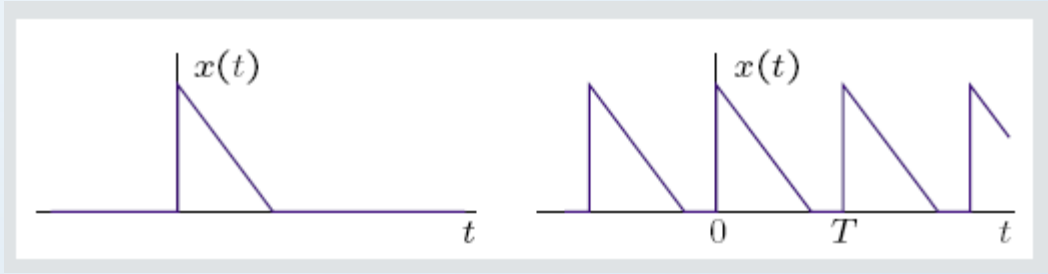
- CT signal is take on real or complex values as a function of an independent variable that ranges over the real numbers and are denoted as x(t).
- DT signals take on real or complex values as a function of an independent variable that ranges over the integers and are denoted as x[n].
- Note the subtle use of parentheses and square brackets to distinguish between CT and DT signals.

A continuous-time signal:

A discrete-time signal is often obtained by sampling a continuous time signal:
T_s = sampling period

Periodic vs Non Periodic Signals

- Periodic signals have the property that x(t + T) = x(t) for all t.
- The smallest value of T that satisfies the definition is called the period.
- Shown below are an non-periodic signal (left) and a periodic signal (right).

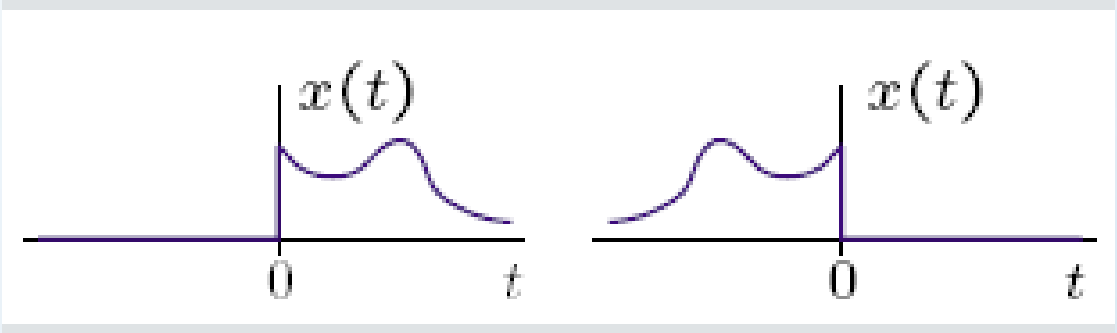


If $x(t)$ is periodic, then

$$x(t) = x(t + nT_0), \quad for\ T_0 \neq 0\ and\ \forall\ integers\ n$$

Casual vs. Non Casual

- A causal signal is zero for t < 0 and an non casual signal is zero for t > 0



Deterministic & Random Signals

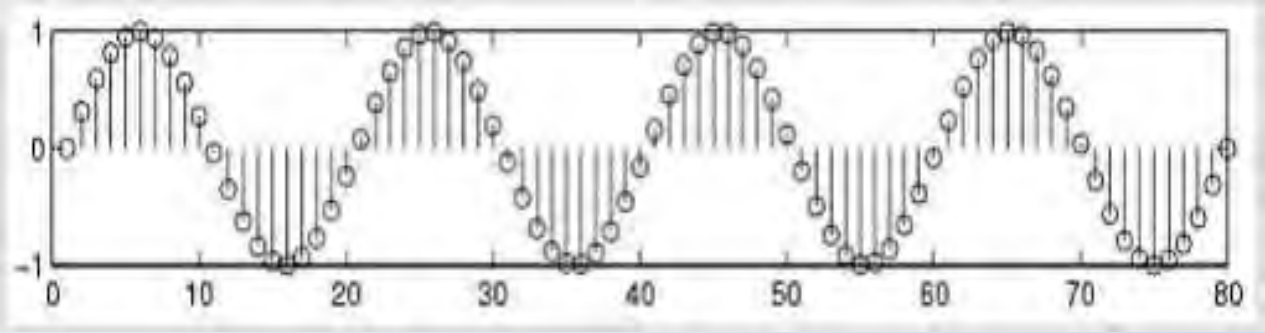
Deterministic signals :

- Behavior of these signals is predictable w.r.t time
- There is no uncertainty with respect to its value at any time.
- These signals can be expressed mathematically.

$$x(t) = \cos(wt + \theta)$$

$$u(t) = \begin{cases} 1, & t > 0 \\ 0, & t < 0 \end{cases}$$

- For example x(t) = sin(3t) is deterministic signal.

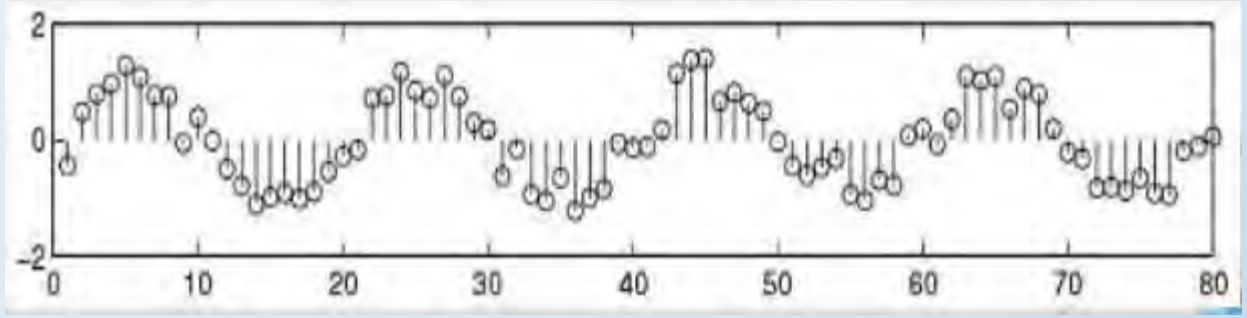


Random Signals:

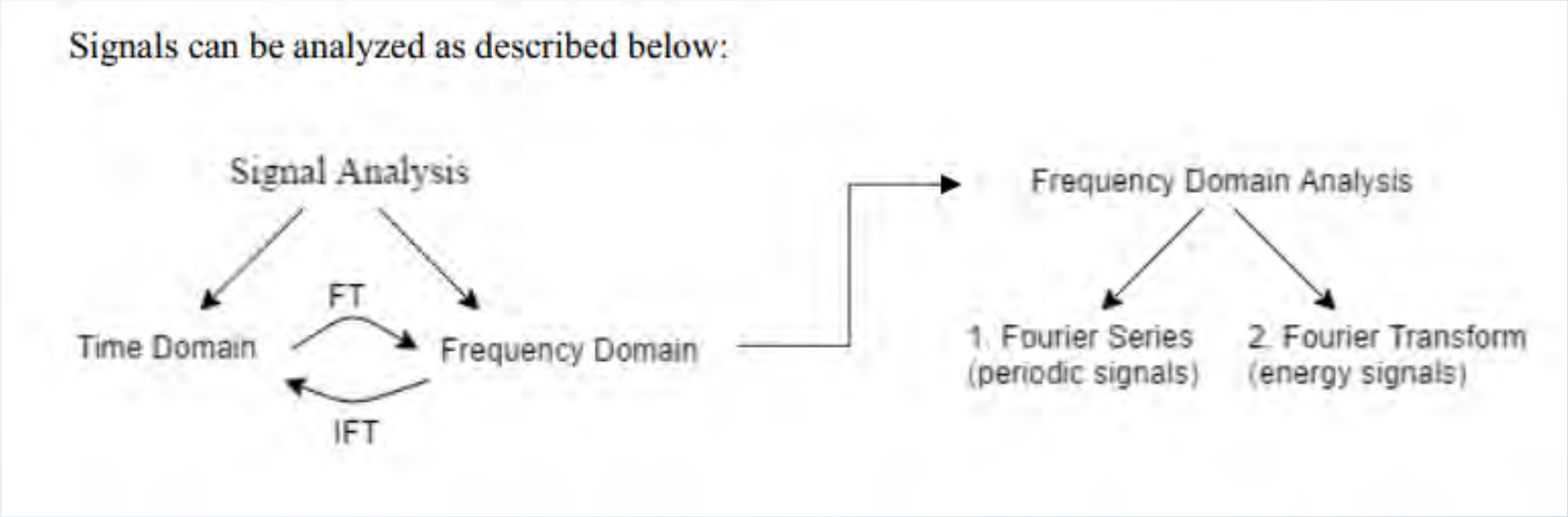
- Behavior of these signals is random i.e. not predictable w.r.t time.
- There is an uncertainty with respect to its value at any time.
- These signals can be analyzed for their characteristics such as expected value, variance, probability density function, etc. As an example, the normal distribution is given as:

$$X \sim N(\mu, \sigma_x^2) = f_x(x) = \frac{1}{\sqrt{2\pi\sigma_x^2}} e^{\frac{(x-\mu)^2}{2\sigma_x^2}}$$

- For example: Thermal Noise generated is non deterministic signal.



Signal Analysis

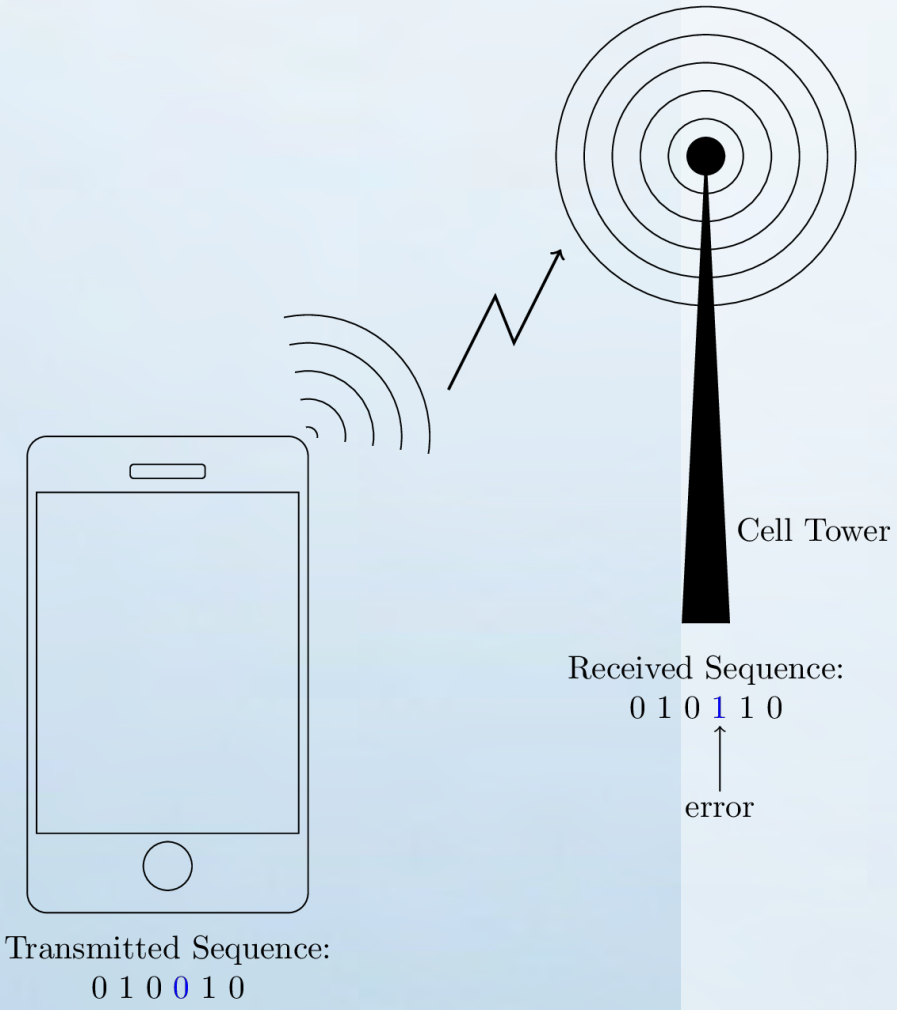


- **What Is Probability?**
- Randomness and uncertainty exist in our daily lives as well as in every discipline in science, engineering, and technology.
- Probability theory is a mathematical framework that allows us to describe and analyze random phenomena in the world around us.
- By random phenomena, we mean events or experiments whose outcomes we can't predict with certainty.
- Example:
- Let's consider a couple of specific applications of probability in order to get some intuition. First, let's think more carefully about what we mean by the terms "randomness" and "probability" in the context of one of the simplest possible random experiments: flipping a fair coin.

- There are two common interpretations of the word "probability.
- " One is in terms of relative frequency. In other words, if we flip the coin a very large number of times, it will come up heads about 1/2 of the time.
- Second interpretation of probability is that it is a quantification of our degree of subjective personal belief that something will happen. example: predicting the weather. When we think about the chances that it will rain today, we consider things like whether there are clouds in the sky and the humidity. However, the beliefs that we form based on these factors may vary from person to person - different people may make different estimates of the probability that it will rain.

Example: Communication Systems

Communication systems play a central role in our lives.



For example, when you talk on the phone, what you say is converted to a sequence of 0's or 1s called *information bits*. These information bits are then transmitted by your cell phone antenna to a nearby cell tower as shown in Figure

Review of Set Theory

Probability theory uses the language of sets.

A **set** is a collection of some items (elements). To define a set we can simply list all the elements in curly brackets,

For example to define a set A that consists of the two elements ♣ and ◊, we write A={♣,◊}

To say that an element does not belong to a set, we use ∉. For example, we may write ♡∉A♡∉A.

A **set** is a collection of things (elements).

Note that ordering does not matter

- The set of natural numbers, $N=\{1,2,3,\cdots\}$
- The set of integers, $Z=\{\cdots,-3,-2,-1,0,1,2,3,\cdots\}$
- The set of rational numbers Q.
- The set of real numbers R.
- Closed intervals on the real line. For example, $[2,3]$ is the set of all real numbers x such that $2\leq x\leq 3$
- Open intervals on the real line. For example $(-1,3)$ is the set of all real numbers x such that $-1<x<3$.
- Similarly, $[1,2)$ is the set of all real numbers x such that $1\leq x<2$.
- The set of complex numbers C is the set of numbers in the form of $a+bi$, where $a,b\in R$,.

We can also define a set by mathematically stating the properties satisfied by the elements in the set. In particular, we may write

$$A=\{x|x \text{ satisfies some property}\}$$

or

$$A=\{x:x \text{ satisfies some property}\}$$

Set A is a **subset** of set B if every element of A is also an element of B. We write $A \subset B$, where "$\subset$" indicates "subset."

Equivalently, we say B is a **superset** of A, or $B \supset A$.

Example

Here are some examples of sets and their subsets:

- If $E=\{1,4\}$ and $C=\{1,4,9\}$, then $E \subset C$.
- $N \subset Z$.
- $Q \subset R$

Two sets are equal if they have the exact same elements. Thus, $A=B$ if and only if $A \subset B$ and $B \subset A$.

For example, $\{1,2,3\}=\{3,2,1\}$,and $\{a,a,b\}=\{a,b\}$.

The set with no elements, i.e., $\emptyset=\{\}$ is the **null set** or the **empty set**.

For any set A, $\emptyset \subset A$.

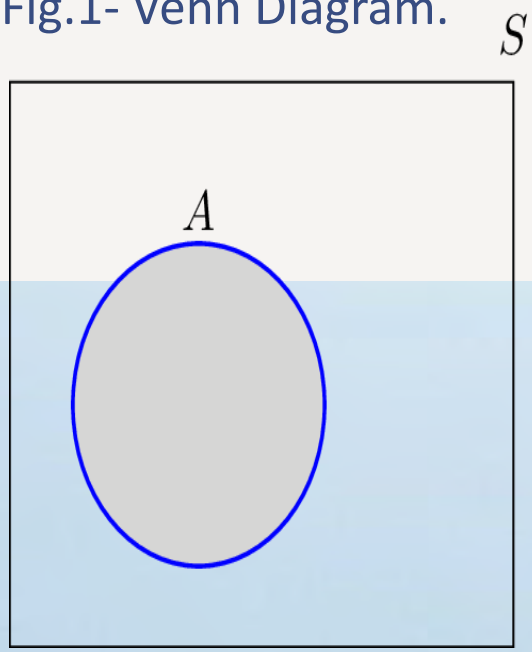
The **universal set** is the set of all things. Thus every set A is a subset of the universal set. We often denote the universal set by S (As we will see, in the language of probability theory, the universal set is called the *sample space*.)

For example, if we are discussing rolling of a die, our universal set may be defined as $S=\{1,2,3,4,5,6\}$

Venn Diagrams

Venn diagrams are very useful in visualizing relation between sets. In a **Venn diagram** any set is depicted by a closed region. Figure 1 shows an example of a Venn diagram. In this figure, the big rectangle shows the universal set S. The shaded area shows another set A.

Fig.1- Venn Diagram.



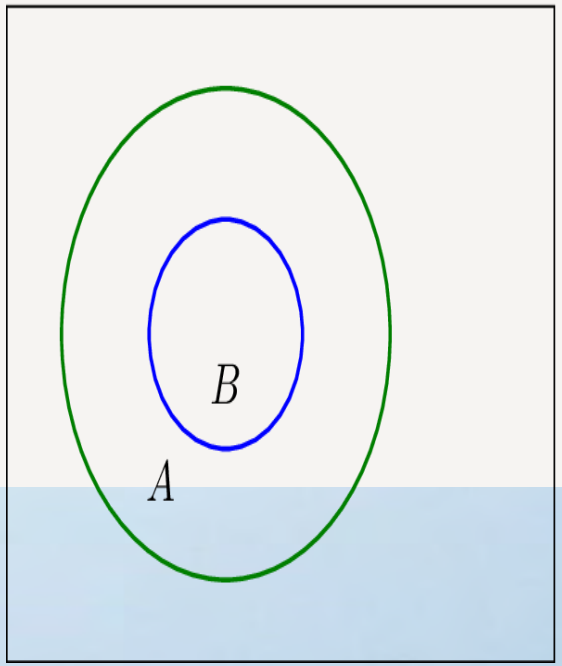


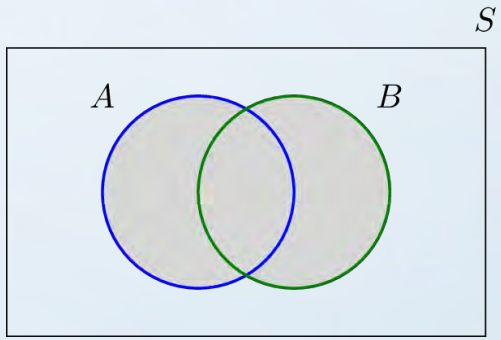
Fig.2 - Venn Diagram for two sets A and B, where $B \subset A$.

Set Operations

Union

The **union** of two sets is a set containing all elements that are in A or in B (possibly both). For example, $\{1,2\}\cup\{2,3\}=\{1,2,3\}$. Thus, we can write $x\in(A\cup B)$ if and only if $(x\in A)$or $(x\in B)$.

Example: $A\cup B=B\cup A$. As shown in the figure the union of sets A and B is shown by the shaded area in the Venn diagram.



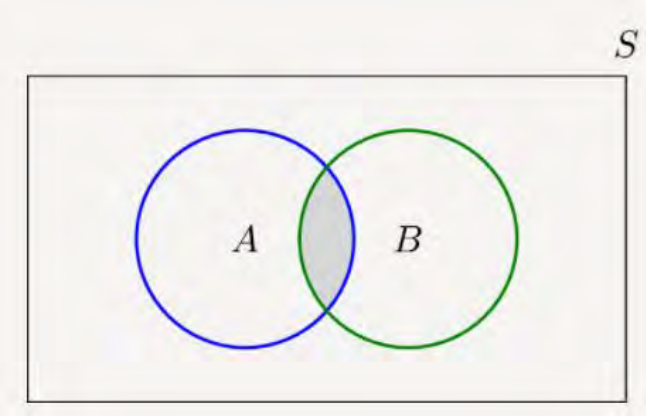
Similarly we can define the union of three or more sets. In particular, if $A_1,A_2,A_3,\cdots,A_n$ are n sets, their union $A_1\cup A_2\cup A_3\cdots\cup A_n$ is a set containing all elements that are in at least one of the sets. We can write this union more compactly by

$$\bigcup_{i=1}^{n} A_i.$$

Set Operations

Intersection

The **intersection** of two sets A and B, denoted by A∩B, consists of all elements that are both in A and B. For example, {1,2}∩{2,3}={2}. As shown in the figure intersection of sets A and B is shown by the shaded area using a Venn diagram.

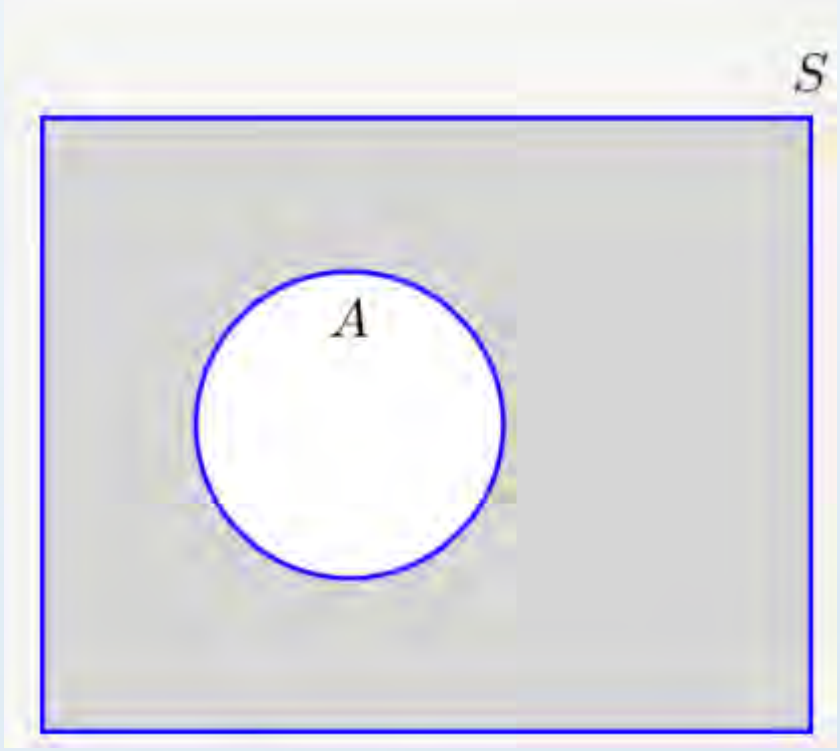


More generally, for sets $A_1, A_2, A_3, \cdots$, their intersection $\bigcap_i A_i$ is defined as the set consisting of the elements that are in all A_i's.

Set Operations

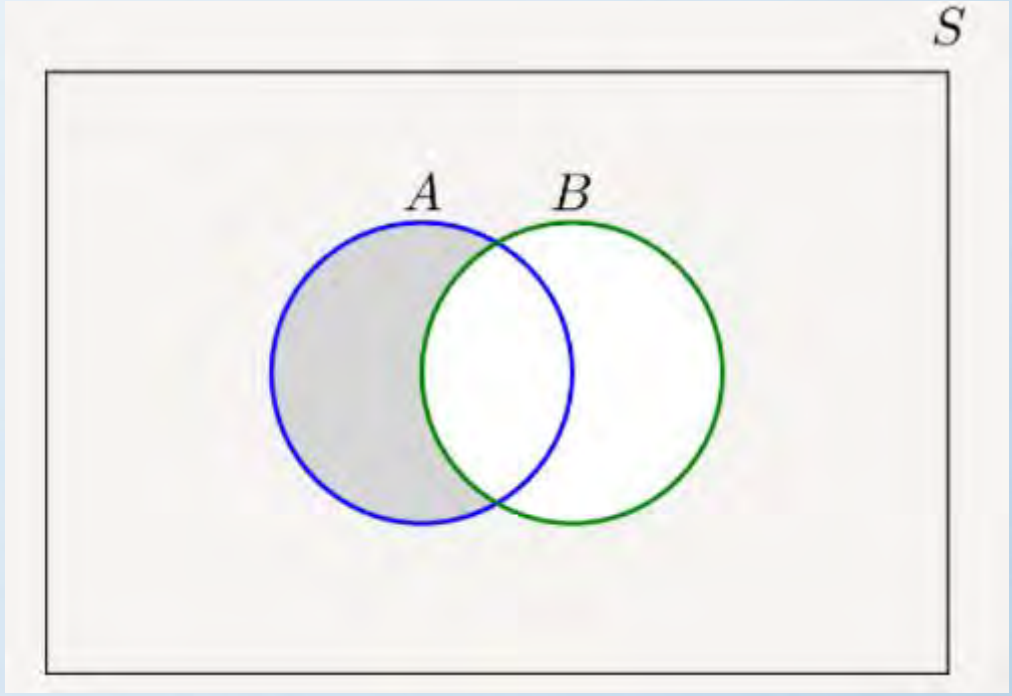
Complement

The **complement** of a set A, denoted $\overline{A}$ by A^c or , is the set of all elements that are in the universal set S but are not in A. As shown in the below figure, $\overline{A}$ is shown by the shaded area using a Venn diagram.



Difference

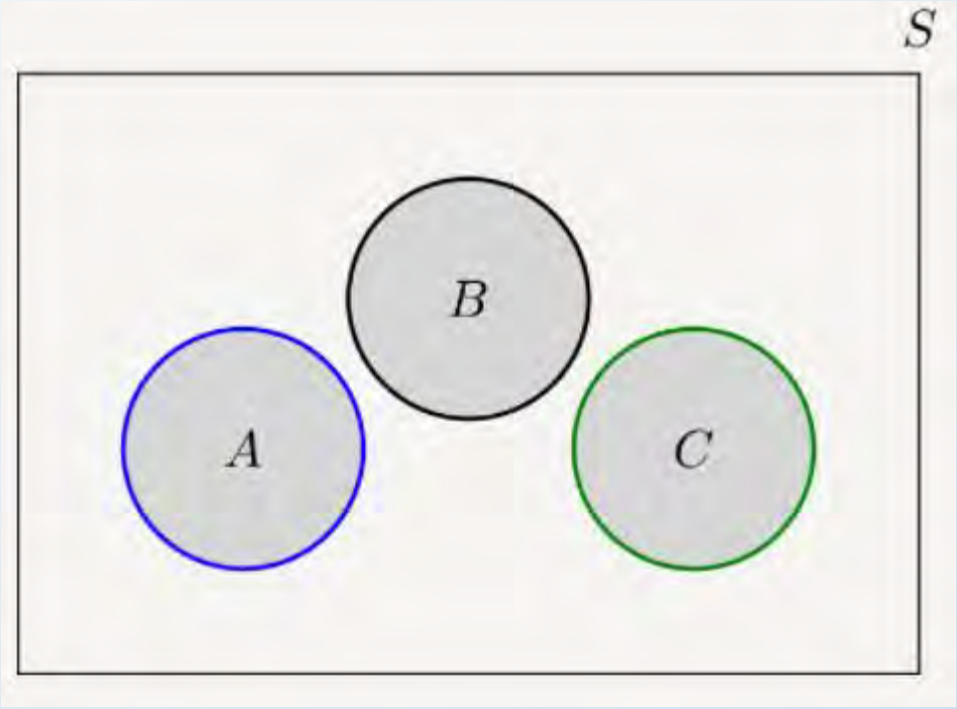
The **difference (subtraction)** is defined as follows. The set $A - B$ consists of elements that are in A but not in B. For example if $A = \{1,2,3\}$ and $B = \{3,5\}$, then $A - B = \{1,2\}$. In Figure $A - B$ is shown by the shaded area using a Venn diagram. Note that $A - B = A \cap B^c$.



Set Operation

Mutually Exclusive/ Disjoin

Two sets A and B are **mutually exclusive** or **disjoint** if they do not have any shared elements; i.e., their intersection is the empty set, $A \cap B = \emptyset$. More generally, several sets are called disjoint if they are pairwise disjoint, i.e., no two of them share a common elements. Figure shows three disjoint sets.



De Morgan's Law

For any sets $A_1, A_2, \cdots, A_n$, we have

- $(A_1 \cup A_2 \cup A_3 \cup \cdots A_n)^c = A_1^c \cap A_2^c \cap A_3^c \cdots \cap A_n^c$;
- $(A_1 \cap A_2 \cap A_3 \cap \cdots A_n)^c = A_1^c \cup A_2^c \cup A_3^c \cdots \cup A_n^c$.

Distributive Law

For any sets A, B, and C we have

- $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$;
- $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.

Example

If the universal set is given by $S = \{1, 2, 3, 4, 5, 6\}$, and $A = \{1, 2\}$, $B = \{2, 4, 5\}, C = \{1, 5, 6\}$ are three sets, find the following sets:

a. $A \cup B$
b. $A \cap B$
c. $\overline{A}$
d. $\overline{B}$
e. Check De Morgan's law by finding $(A \cup B)^c$ and $A^c \cap B^c$.
f. Check the distributive law by finding $A \cap (B \cup C)$ and $(A \cap B) \cup (A \cap C)$.

a. $A \cup B = \{1, 2, 4, 5\}$.
b. $A \cap B = \{2\}$.
c. $\overline{A} = \{3, 4, 5, 6\}$ ($\overline{A}$ consists of elements that are in S but not in A).
d. $\overline{B} = \{1, 3, 6\}$.
e. We have

$$(A \cup B)^c = \{1, 2, 4, 5\}^c = \{3, 6\},$$

which is the same as

$$A^c \cap B^c = \{3, 4, 5, 6\} \cap \{1, 3, 6\} = \{3, 6\}.$$

f. We have

$$A \cap (B \cup C) = \{1, 2\} \cap \{1, 2, 4, 5, 6\} = \{1, 2\}$$

which is the same as

$$(A \cap B) \cup (A \cap C) = \{2\} \cup \{1\} = \{1, 2\}.$$

Cartesian Product

A **Cartesian product** of two sets A and B, written as $A \times B$, is the set containing **ordered** pairs from A and B. That is, if $C = A \times B$, then each element of C is of the form (x, y), where $x \in A$ and $y \in B$:

$$A \times B = \{(x, y) | x \in A \text{ and } y \in B\}.$$

For example, if $A = \{1, 2, 3\}$ and $B = \{H, T\}$, then

$$A \times B = \{(1, H), (1, T), (2, H), (2, T), (3, H), (3, T)\}.$$

Note that here the pairs are ordered, so for example, $(1, H) \neq (H, 1)$. Thus $A \times B$ is **not** the same as $B \times A$.

If you have two finite sets A and B, where A has M elements and B has N elements, then $A \times B$ has $M \times N$ elements. This rule is called the **multiplication principle** and is very useful in counting the numbers of elements in sets. The number of elements in a set is denoted by $|A|$, so here we write $|A| = M, |B| = N$, and $|A \times B| = MN$. In the above example, $|A| = 3, |B| = 2$, thus $|A \times B| = 3 \times 2 = 6$. We can similarly define the Cartesian product of n sets $A_1, A_2, \cdots, A_n$ as

$$A_1 \times A_2 \times A_3 \times \cdots \times A_n = \{(x_1, x_2, \cdots, x_n) | x_1 \in A_1 \text{ and } x_2 \in A_2 \text{ and } \cdots x_n \in A_n\}.$$

The multiplication principle states that for finite sets $A_1, A_2, \cdots, A_n$, if

$$|A_1| = M_1, |A_2| = M_2, \cdots, |A_n| = M_n,$$

then

$$|A_1 \times A_2 \times A_3 \times \cdots \times A_n| = M_1 \times M_2 \times M_3 \times \cdots \times M_n.$$

An important example of sets obtained using a Cartesian product is $\mathbb{R}^n$, where n is a natural number. For $n = 2$, we have

$$\begin{aligned} \mathbb{R}^2 &= \mathbb{R} \times \mathbb{R} \\ &= \{(x, y) | x \in \mathbb{R}, y \in \mathbb{R}\}. \end{aligned}$$

Thus, $\mathbb{R}^2$ is the set consisting of all points in the two-dimensional plane. Similarly, $\mathbb{R}^3 = \mathbb{R} \times \mathbb{R} \times \mathbb{R}$ and so on.

Cardinality: Countable and Uncountable Sets

- **Cardinality** of a set, which is basically the size of the set. The cardinality of a set is denoted by $|A|$.
- We will first discuss cardinality for finite sets and then talk about infinite sets.

Finite Set

Consider a set A. If A has only a finite number of elements, its cardinality is simply the number of elements in A. For example, if $A=\{2,4,6,8,10\}$, then $|A|=5$.

Inclusion-Exclusion Principal

Inclusion-exclusion principle:

1. $|A \cup B| = |A| + |B| - |A \cap B|$,
2. $|A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |A \cap C| - |B \cap C| + |A \cap B \cap C|$.

Generally, for n finite sets $A_1, A_2, A_3, \cdots, A_n$, we can write

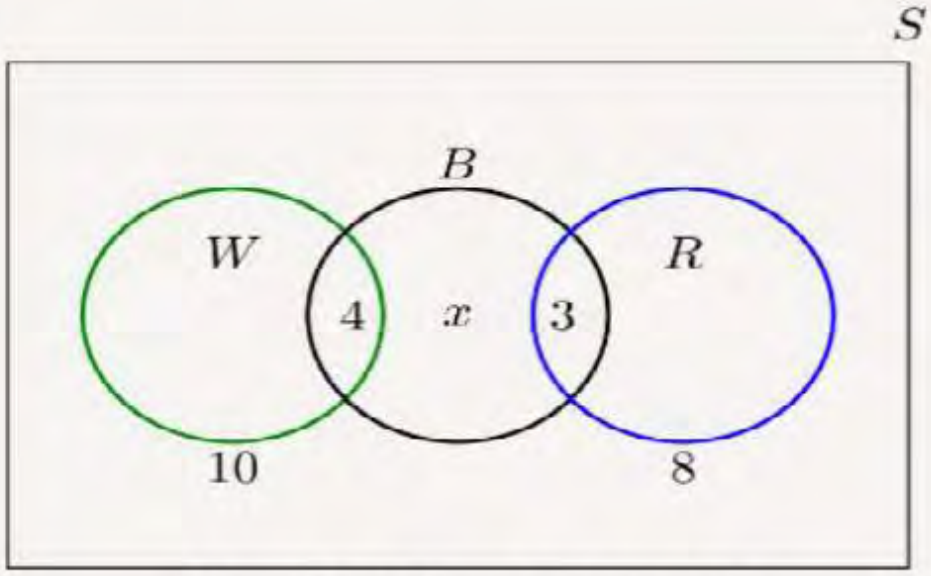
$$\left|\bigcup_{i=1}^{n} A_i\right| = \sum_{i=1}^{n} |A_i| - \sum_{i<j} |A_i \cap A_j|$$

$$+ \sum_{i<j<k} |A_i \cap A_j \cap A_k| - \cdots + (-1)^{n+1} |A_1 \cap \cdots \cap A_n|.$$

In a party,

- there are 10 people with white shirts and 8 people with red shirts;
- 4 people have black shoes and white shirts;
- 3 people have black shoes and red shirts;
- the total number of people with white or red shirts or black shoes is 21.

How many people have black shoes?



$$21 = 10 + 8 + x \Rightarrow x = 3$$
$$\Rightarrow |B| = 4 + x + 3 = 10$$

Let W, R, and B, be the number of people with white shirts, red shirts, and black shoes respectively. Then, here is the summary of the available information:

$$|W| = 10$$

$$|R| = 8$$

$$|W \cap B| = 4$$

$$|R \cap B| = 3$$

$$|W \cup B \cup R| = 21.$$

Also, it is reasonable to assume that W and R are disjoint, $|W \cap R| = 0$. Thus by applying the inclusion-exclusion principle we obtain

$$\begin{aligned} |W \cup R \cup B| &= 21 \\ &= |W| + |R| + |B| - |W \cap R| - |W \cap B| - |R \cap B| + |W \cap R \cap B| \\ &= 10 + 8 + |B| - 0 - 4 - 3 + 0. \end{aligned}$$

Thus

$$|B| = 10.$$

Infinite Sets

There are two types of infinite sets, where one type is significantly "larger" than the other. In particular, one type is called **countable**, while the other is called **uncountable**.

Sets such as $\mathbb{N}$ and $\mathbb{Z}$ are called countable, but "bigger" sets such as $\mathbb{R}$ are called uncountable.

The difference between the two types is that you can list the elements of a countable set A, i.e., you can write $A=\{a_1,a_2,\cdots\}$, but you cannot list the elements in an uncountable set. For example, you can write

- $\mathbb{N}=\{1,2,3,\cdots\}$,
- $\mathbb{Z}=\{0,1,-1,2,-2,3,-3,\cdots\}$.

The fact that you can list the elements of a countably infinite set means that the set can be put in one-to-one correspondence with natural numbers $\mathbb{N}$.

On the other hand, you cannot list the elements in $\mathbb{R}$, so it is an uncountable set.

Definition

Set A is called countable if one of the following is true

a. if it is a finite set, $|A| < \infty$; or
b. it can be put in one-to-one correspondence with natural numbers $\mathbb{N}$, in which case the set is said to be countably infinite.
A set is called uncountable if it is not countable.

- $\mathbb{N}, \mathbb{Z}, \mathbb{Q}$, and any of their subsets are countable.
- Any set containing an interval on the real line such as $[a, b], (a, b], [a, b)$, or (a, b), where $a < b$ is uncountable.

Therom 1

Any subset of a countable set is countable.
Any superset of an uncountable set is uncountable.

Proof
The intuition behind this theorem is the following: If a set is countable, then any "smaller" set should also be countable, so a subset of a countable set should be countable as well. To provide a proof, we can argue in the following way.

Let A be a countable set and $B \subset A$. If A is a finite set, then $|B| \leq |A| < \infty$, thus B is countable. If A is countably infinite, then we can list the elements in A, then by removing the elements in the list that are not in B, we can obtain a list for B, thus B is countable.

The second part of the theorem can be proved using the first part. Assume B is uncountable. If $B \subset A$ and A is countable, by the first part of the theorem B is also a countable set which is a contradiction.

Therom 2

If $A_1, A_2, \cdots$ is a list of countable sets, then the set $\bigcup_i A_i = A_1 \cup A_2 \cup A_3 \cdots$ is also countable.

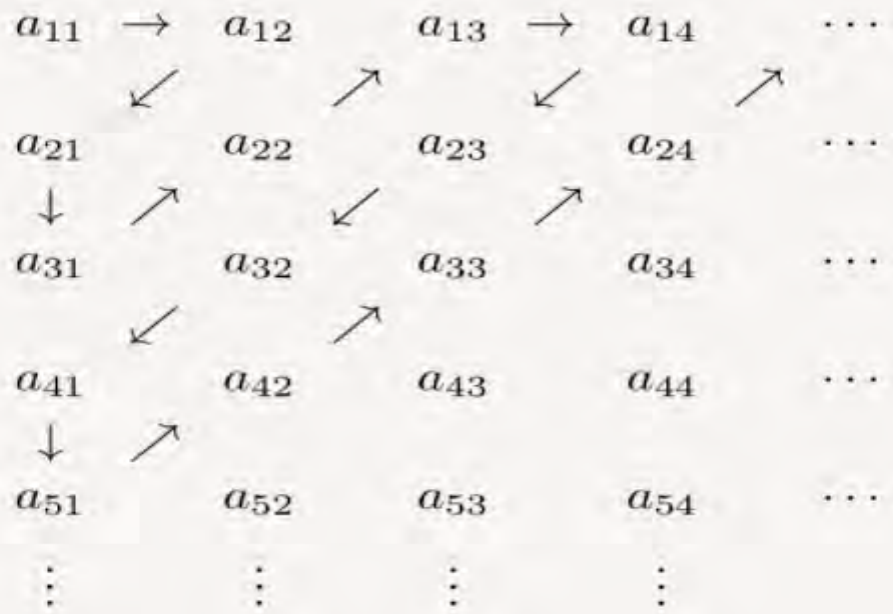
Proof

It suffices to create a list of elements in $\bigcup_i A_i$. Since each A_i is countable we can list its elements: $A_i = \{a_{i1}, a_{i2}, \cdots\}$. Thus, we have

- $A_1 = \{a_{11}, a_{12}, \cdots\}$,
- $A_2 = \{a_{21}, a_{22}, \cdots\}$,
- $A_3 = \{a_{31}, a_{32}, \cdots\}$,
- $\cdots$

Now we need to make a list that contains all the above lists. This can be done in different ways. One way to do this is to use the ordering shown in Figure 1.12 to make a list. Here, we can write

$$\bigcup_i A_i = \{a_{11}, a_{12}, a_{21}, a_{31}, a_{22}, a_{13}, a_{14}, \cdots\} \tag{1.1}$$



We have been able to create a list that contains all the elements in $\bigcup_i A_i$, so this set is countable.

FUNCTIONS

We often need the concept of functions in probability. A function f is a rule that takes an input from a specific set, called the **domain**, and produces an output from another set, called **co-domain**. Thus, a function *maps* elements from the domain set to elements in the co-domain with the property that each input is mapped to exactly one output. For a function f, if x is an element in the domain, then the function value (the output of the function) is shown by $f(x)$. If A is the domain and B is the co-domain for the function f, we use the following notation:

$$f : A \to B.$$

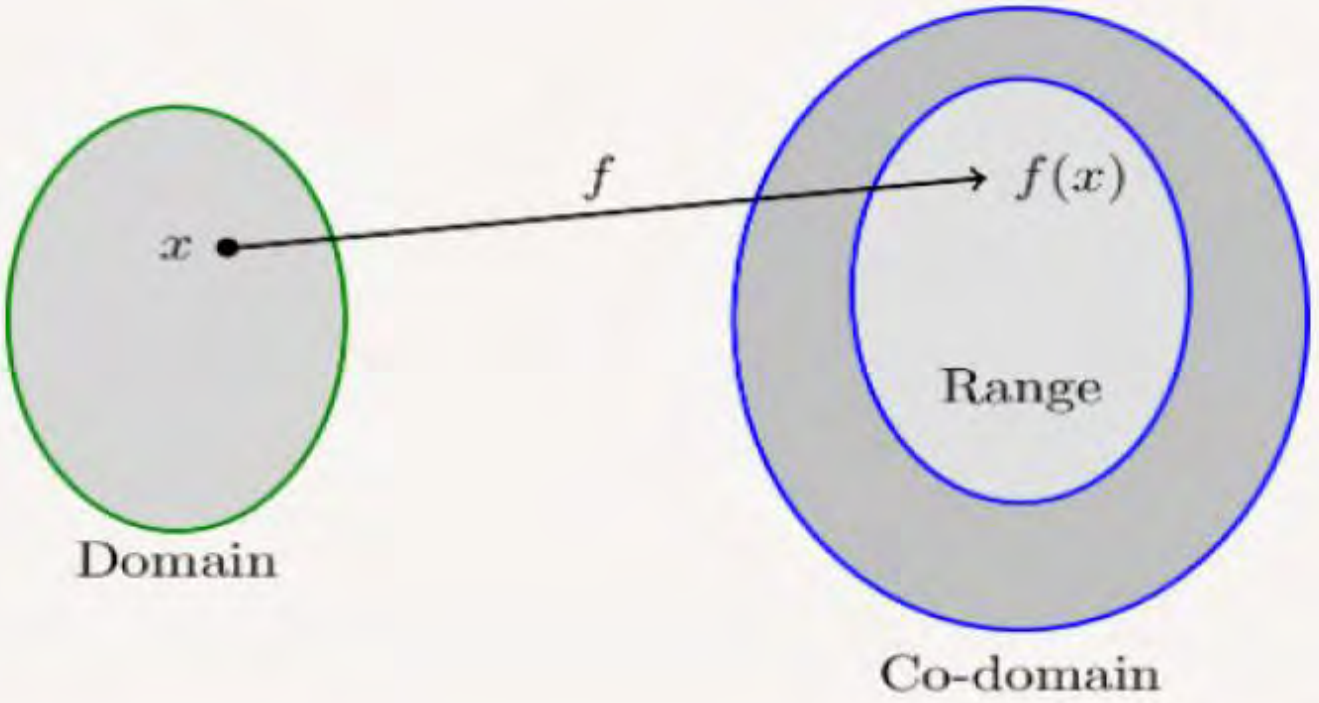
Example

- Consider the function $f : \mathbb{R} \to \mathbb{R}$, defined as $f(x) = x^2$. This function takes any real number x and outputs x^2. For example, $f(2) = 4$.
- Consider the function $g : \{H, T\} \to \{0, 1\}$, defined as $g(H) = 0$ and $g(T) = 1$. This function can only take two possible inputs H or T, where H is mapped to 0 and T is mapped to 1.

The output of a function $f : A \to B$ always belongs to the co-domain B. However, not all values in the co-domain are always covered by the function. In the above example, $f : \mathbb{R} \to \mathbb{R}$, the function value is always a positive number $f(x) = x^2 \geq 0$. We define the **range** of a function as the set containing all the possible values of $f(x)$. Thus, the range of a function is always a subset of its co-domain. For the above function $f(x) = x^2$, the range of f is given by

$$\textbf{Range}(f) = \mathbb{R}^+ = \{x \in \mathbb{R} | x \geq 0\}.$$

Figure 1.14 pictorially shows a function, its domain, co-domain, and range. The figure shows that an element x in the domain is mapped to $f(x)$ in the range.



Random Experiments and Probabilities

Random Experiments

- In particular, a random experiment is a process by which we observe something uncertain. After the experiment, the result of the random experiment is known.
- An **outcome** is a result of a random experiment.
- The set of all possible outcomes is called the **sample space**. Thus in the context of a random experiment, the sample space is our *universal set*.
- Here are some examples of random experiments and their sample spaces:

  - Random experiment: toss a coin; sample space: $S = \{heads, tails\}$ or as we usually write it, $\{H, T\}$.
  - Random experiment: roll a die; sample space: $S = \{1, 2, 3, 4, 5, 6\}$.
  - Random experiment: observe the number of iPhones sold by an Apple store in Boston in 2015; sample space: $S = \{0, 1, 2, 3, \cdots\}$.
  - Random experiment: observe the number of goals in a soccer match; sample space: $S = \{0, 1, 2, 3, \cdots\}$.

Random Experiments

- When we repeat a random experiment several times, we call each one of them a **trial**.
- Thus, a trial is a particular performance of a random experiment.
- In the example of tossing a coin, each trial will result in either heads or tails.
- Note that the sample space is defined based on how you define your random experiment. For example,

We toss a coin three times and observe the sequence of heads/tails. The sample space here may be defined as

$$S = \{(H,H,H),(H,H,T),(H,T,H),(T,H,H),(H,T,T),(T,H,T),(T,T,H),(T,T,T)\}.$$

Events.

- For example, suppose that we would like to know the probability that the outcome of rolling a fair die is an even number.
- In this case, our event is the set $E=\{2,4,6\}$. If the result of our random experiment belongs to the set E, we say that the event E has occurred.
- Thus an event is a collection of possible outcomes. In other words, an event is a subset of the sample space to which we assign a probability.

Outcome: A result of a random experiment.
Sample Space: The set of all possible outcomes.
Event: A subset of the sample space.

Probability

- A **probability** measure P(A) to an event A. This is a value between 0 and 1 that shows how likely the event is.
- If P(A) is close to 0, it is very unlikely that the event A occurs. On the other hand, if P(A) is close to 1, A is very likely to occur.
- The main subject of probability theory is to develop tools and techniques to calculate probabilities of different events.
- Probability theory is based on some axioms that act as the foundation for the theory.

Axioms of Probability:

- Axiom 1: For any event A, $P(A) \geq 0$.
- Axiom 2: Probability of the sample space S is $P(S) = 1$.
- Axiom 3: If $A_1, A_2, A_3, \cdots$ are disjoint events, then $P(A_1 \cup A_2 \cup A_3 \cdots) = P(A_1) + P(A_2) + P(A_3) + \cdots$

- The first axiom states that probability cannot be negative. The smallest value for $P(A)$ is zero and if $P(A)=0$, then the event A will never happen.
- The second axiom states that the probability of the whole sample space is equal to one, i.e., 100 percent. The reason for this is that the sample space S contains all possible outcomes of our random experiment. Thus, the outcome of each trial always belongs to S, i.e., the event S always occurs and $P(S)=1$.
- The third axiom, the basic idea is that if some events are disjoint (i.e., there is no overlap between them), then the probability of their union must be the summations of their probabilities.
- Another way to think about this is to imagine the probability of a set as the area of that set in the Venn diagram.

In a presidential election, there are four candidates. Call them A, B, C, and D. Based on our polling analysis, we estimate that A has a 20 percent chance of winning the election, while B has a 40 percent chance of winning. What is the probability that A **or** B win the election?

Notice that the events that $\{\text{A wins}\}$, $\{\text{B wins}\}$, $\{\text{C wins}\}$, and $\{\text{D wins}\}$ are disjoint since more than one of them cannot occur at the same time. For example, if A wins, then B cannot win. From the third axiom of probability, the probability of the union of two disjoint events is the summation of individual probabilities. Therefore,

$$\begin{aligned} P(\text{A wins or B wins}) &= P(\{\text{A wins}\} \cup \{\text{B wins}\}) \\ &= P(\{\text{A wins}\}) + P(\{\text{B wins}\}) \\ &= 0.2 + 0.4 \\ &= 0.6 \end{aligned}$$

As we have seen, when working with events, *intersection* means *"and"*, and *union* means *"or"*. The probability of intersection of A and B, $P(A \cap B)$, is sometimes shown by $P(A, B)$ or $P(AB)$.

Notation:

- $P(A \cap B) = P(A \text{ and } B) = P(A, B)$,
- $P(A \cup B) = P(A \text{ or } B)$.

Lecture 2

Finding Probabilities

Suppose that we are given a random experiment with a sample space S. To find the probability of an event, there are usually two steps:

- first, we use the specific information that we have about the random experiment.
- Second, we use the probability axioms.

Let's look at an example.

- You roll a fair die. What is the probability of E={1,5}E={1,5}?

Let's first use the specific information that we have about the random experiment. The problem states that the die is fair, which means that all six possible outcomes are equally likely, i.e.,

$$P(\{1\}) = P(\{2\}) = \cdots = P(\{6\}).$$

Now we can use the axioms of probability. In particular, since the events $\{1\}, \{2\}, \cdots, \{6\}$ are disjoint we can write

$$\begin{aligned} 1 &= P(S) \\ &= P\Big(\{1\} \cup \{2\} \cup \cdots \cup \{6\}\Big) \\ &= P(\{1\}) + P(\{2\}) + \cdots + P(\{6\}) \\ &= 6P(\{1\}). \end{aligned}$$

Thus,

$$P(\{1\}) = P(\{2\}) = \cdots = P(\{6\}) = \frac{1}{6}.$$

Again since $\{1\}$ and $\{5\}$ are disjoint, we have

$$P(E) = P(\{1,5\}) = P(\{1\}) + P(\{5\}) = \frac{2}{6} = \frac{1}{3}.$$

Generally for n events $A_1, A_2, \cdots, A_n$, we have

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) - \sum_{i<j} P(A_i \cap A_j)$$
$$+ \sum_{i<j<k} P(A_i \cap A_j \cap A_k) - \cdots + (-1)^{n-1} P\left(\bigcap_{i=1}^{n} A_i\right)$$

Discrete Probability Models

Consider a sample space S. If S is a *countable* set, this refers to a **discrete** probability model. In this case, since S is countable, we can list all the elements in S:

$$S = \{s_1, s_2, s_3, \cdots\}.$$

If $A \subset S$ is an event, then A is also countable, and by the third axiom of probability we can write

$$P(A) = P(\bigcup_{s_j \in A} \{s_j\}) = \sum_{s_j \in A} P(s_j).$$

Thus in a countable sample space, to find probability of an event, all we need to do is sum the probability of individual elements in that set.

Example

I play a gambling game in which I will win $k-2$ dollars with probability $\frac{1}{2^k}$ for any $k \in \mathbb{N}$, that is,

- with probability $\frac{1}{2}$, I lose 1 dollar;
- with probability $\frac{1}{4}$, I win 0 dollar;
- with probability $\frac{1}{8}$, I win 1 dollar;
- with probability $\frac{1}{16}$, I win 2 dollars;
- with probability $\frac{1}{32}$, I win 3 dollars;
- ...

What is the probability that I win more than or equal to 1 dollar and less than 4 dollars? What is the probability that I win more than 2 dollars?

In this problem, the random experiment is the gambling game and the outcomes are the amount in dollars that I win (lose). Thus we may write

$$S = \{-1, 0, 1, 2, 3, 4, 5, \cdots\}.$$

As we see this is an infinite but countable set. The problem also states that

$$P(k) = P(\{k\}) = \frac{1}{2^{k+2}} \text{ for } k \in S.$$

First, let's check that this is a valid probability measure. To do so, we should check if all probabilities add up to one, i.e., $P(S) = 1$. We have

$$\begin{aligned} P(S) &= \sum_{k=-1}^{\infty} P(k) \\ &= \sum_{k=-1}^{\infty} \frac{1}{2^{k+2}} \\ &= \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots \qquad \text{(geometric sum)} \\ &= 1. \end{aligned}$$

Now let's solve the problem. Let's define A as the event that I win more than or equal to 1 dollar and less than 4 dollars, and B as the event that I win more than 2 dollars. Thus,

$$A = \{1, 2, 3\}, B = \{3, 4, 5, \cdots\}.$$

Then

$$\begin{aligned} P(A) &= P(1) + P(2) + P(3) \\ &= \frac{1}{8} + \frac{1}{16} + \frac{1}{32} \\ &= \frac{7}{32} \\ &\approx 0.219 \end{aligned}$$

Similarly,

$$\begin{aligned} P(B) &= P(3) + P(4) + P(5) + P(6) + \cdots \\ &= \frac{1}{32} + \frac{1}{64} + \frac{1}{128} + \frac{1}{256} + \cdots \qquad \text{(geometric sum)} \\ &= \frac{1}{16} \\ &= 0.0625 \end{aligned}$$

Continuous Probability Models

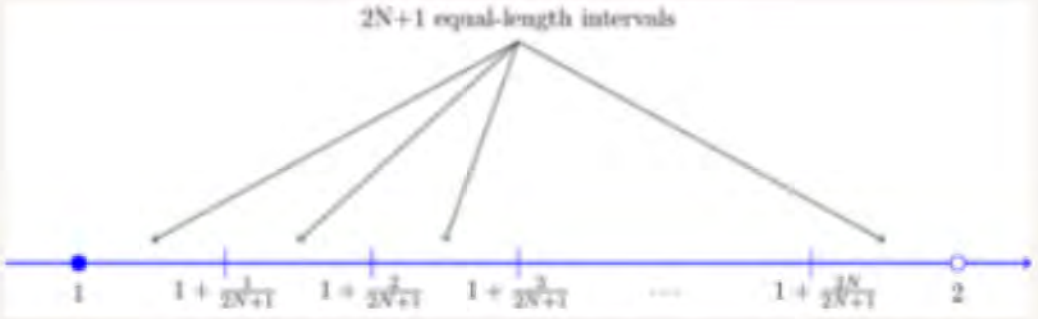
- Consider a scenario where your sample space S is, for example, [0,1]. This is an uncountable set; we cannot list the elements in the set.
- At this time, we have not yet developed the tools needed to deal with continuous probability models, but we can provide some intuition by looking at a simple example.

Your friend tells you that she will stop by your house sometime after or equal to 1 p.m. and before 2 p.m., but she cannot give you any more information as her schedule is quite hectic. Your friend is very dependable, so you are sure that she will stop by your house, but other than that we have no information about the arrival time. Thus, we assume that the arrival time is completely random in the 1 p.m. and 2 p.m. interval. (As we will see, in the language of probability theory, we say that the arrival time is "uniformly" distributed on the $[1,2)$ interval). Let T be the arrival time.

a. What is the sample space S?
b. What is the probability of $P(1.5)$? Why?
c. What is the probability of $T \in [1,1.5)$?
d. For $1 \leq a \leq b \leq 2$, what is $P(a \leq T \leq b) = P([a,b])$?

a. Since any real number in $[1,2)$ is a possible outcome, the sample space is indeed $S=[1,2)$.

b. Now, let's look at $P(1.5)$. A reasonable guess would be $P(1.5)=0$. But can we provide a reason for that? Let us divide the $[1,2)$ interval to $2N+1$ equal-length and disjoint intervals, $[1,1+\frac{1}{2N+1}),[1+\frac{1}{2N+1},1+\frac{2}{2N+1}),\cdots,[1+\frac{N}{2N+1},1+\frac{N+1}{2N+1}),\cdots,[1+\frac{2N}{2N+1},2)$. See Figure [redacted] Here, N could be any positive integer.



The only information that we have is that the arrival time is "uniform" on the $[1,2)$ interval. Therefore, all of the above intervals should have the same probability, and since their union is S we conclude that

$$P\left(\left[1,1+\frac{1}{2N+1}\right)\right)=P\left(\left[1+\frac{1}{2N+1},1+\frac{2}{2N+1}\right)\right)=\cdots$$

$$\cdots=P\left(\left[1+\frac{N}{2N+1},1+\frac{N+1}{2N+1}\right)\right)=\cdots$$

$$\cdots=P\left(\left[1+\frac{2N}{2N+1},2\right)\right)=\frac{1}{2N+1}.$$

In particular, by defining $A_N=\left[1+\frac{N}{2N+1},1+\frac{N+1}{2N+1}\right)$, we conclude that

$$P(A_N)=P\left(\left[1+\frac{N}{2N+1},1+\frac{N+1}{2N+1}\right)\right)=\frac{1}{2N+1}.$$

Now note that for any positive integer N, $1.5\in A_N$. Thus, $\{1.5\}\subset A_N$, so

$$P(1.5)\leq P(A_N)=\frac{1}{2N+1},\qquad \text{for all } N\in\mathbb{N}.$$

Note that as N becomes large, $P(A_N)$ approaches 0. Since $P(1.5)$ cannot be negative, we conclude that $P(1.5)=0$. Similarly, we can argue that $P(x)=0$ for all $x\in[1,2)$.

c. Next, we find $P([1,1.5))$. This is the first half of the entire sample space $S=[1,2)$ and because of uniformity, its probability must be 0.5. In other words,

$$P([1,1.5)) = P([1.5,2)) \qquad \text{(by uniformity)},$$

$$P([1,1.5)) + P([1.5,2)) = P(S) = 1.$$

Thus

$$P([1,1.5)) = P([1.5,2)) = \frac{1}{2}.$$

d. The same uniformity argument suggests that all intervals in $[1,2)$ with the same length must have the same probability. In particular, the probability of an interval is proportional to its length. For example, since

$$[1,1.5) = [1,1.25) \cup [1.25,1.5).$$

Thus, we conclude

$$P\big([1,1.5)\big) = P\big([1,1.25)\big) + P\big([1.25,1.5)\big)$$

$$= 2P\big([1,1.25)\big).$$

And finally, since $P\big([1,2)\big) = 1$, we conclude

$$P([a,b]) = b-a, \qquad \text{for } 1 \le a \le b < 2.$$

Conditional Probability

- In this section, we discuss one of the most fundamental concepts in probability theory.
- For example, suppose that in a certain city, 23% of the days are rainy. Thus, if you pick a random day, the probability that it rains that day is 23%:

$$P(R)=0.23$$, where R is the event that it rains on the randomly chosen.

- Now suppose that I pick a random day, but I also tell you that it is cloudy on the chosen day. Now that you have this extra piece of information, what is the probability that it rains **given that** it is cloudy?
- If C is the event that it is cloudy, then we write this as $P(R|C)$, the *conditional probability of* R *given that* C *has occurred*.
- It is reasonable to assume that in this example, $P(R|C)$ should be larger than the original $P(R)$, which is called the **prior probability** of R.
- But what exactly should $P(R|C)$ be?

Example

If we roll a fair die. Let AA be the event that the outcome is an odd number, i.e., A={1,3,5}. Also let B be the event that the outcome is less than or equal to 3, i.e., B={1,2,3}. What is P(A)? And What is the probability of P(A|B)?

This is a finite sample space, so

$$P(A) = \frac{|A|}{|S|} = \frac{|\{1,3,5\}|}{6} = \frac{1}{2}.$$

Now, let's find the conditional probability of A given that B occurred. If we know B has occurred, the outcome must be among $\{1,2,3\}$. For A to also happen the outcome must be in $A \cap B = \{1,3\}$. Since all die rolls are equally likely, we argue that $P(A|B)$ must be equal to

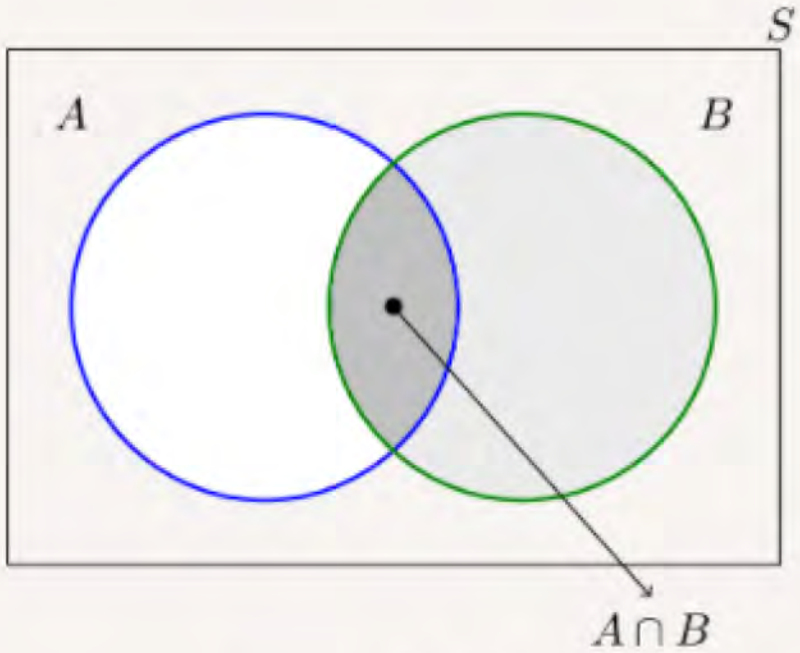
$$P(A|B) = \frac{|A \cap B|}{|B|} = \frac{2}{3}.$$

If A and B are two events in a sample space S, then the **conditional probability of A given B** is defined as

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

Here is the intuition behind the formula. When we know that B has occurred, every outcome that is outside B should be discarded. Thus, *our sample space is reduced to the set* B,
Now the only way that A can happen is when the outcome belongs to the set $A \cap B$. We divide $P(A \cap B)$ by $P(B)$, so that the conditional probability of the new sample space becomes 1, i.e., $P(B|B) = \frac{P(B \cap B)}{P(B)} = 1$.

Note that conditional probability of $P(A|B)$ is undefined when $P(B) = 0$. That is okay because if $P(B) = 0$, it means that the event B never occurs so it does not make sense to talk about the probability of A given B.



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

Venn diagram for conditional probability, $P(A|B)$.

Let's look at some special cases of conditional probability:

- When A and B are disjoint: In this case $A \cap B = \emptyset$, so

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} \\ &= \frac{P(\emptyset)}{P(B)} \\ &= 0. \end{aligned}$$

  This makes sense. In particular, since A and B are disjoint they cannot both occur at the same time. Thus, given that B has occurred, the probability of A must be zero.

- When B is a subset of A: If $B \subset A$, then whenever B happens, A also happens. Thus, given that B occurred, we expect that probability of A be one. In this case $A \cap B = B$, so

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} \\ &= \frac{P(B)}{P(B)} \\ &= 1. \end{aligned}$$

- When A is a subset of B: In this case $A \cap B = A$, so

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} \\ &= \frac{P(A)}{P(B)}. \end{aligned}$$

Example

Consider a family that has two children. We are interested in the children's genders. Our sample space is S={(G,G),(G,B),(B,G),(B,B)}. Also assume that all four possible outcomes are equally likely.

a.What is the probability that both children are girls given that the first child is a girl?

b.If we ask the father: "Do you have at least one daughter?" He responds "Yes!" Given this extra information, what is the probability that both children are girls? In other words, what is the probability that both children are girls given that we know at least one of them is a girl?

Let A be the event that both children are girls, i.e., $A = \{(G,G)\}$. Let B be the event that the first child is a girl, i.e., $B = \{(G,G),(G,B)\}$. Finally, let C be the event that at least one of the children is a girl, i.e., $C = \{(G,G),(G,B),(B,G)\}$. Since the outcomes are equally likely, we can write

$$P(A) = \frac{1}{4},$$

$$P(B) = \frac{2}{4} = \frac{1}{2},$$

$$P(C) = \frac{3}{4}.$$

a. What is the probability that both children are girls given that the first child is a girl? This is $P(A|B)$, thus we can write

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} \\ &= \frac{P(A)}{P(B)} \qquad (\text{since } A \subset B) \\ &= \frac{\frac{1}{4}}{\frac{1}{2}} = \frac{1}{2}. \end{aligned}$$

b. What is the probability that both children are girls given that we know at least one of them is a girl? This is $P(A|C)$, thus we can write

$$\begin{aligned} P(A|C) &= \frac{P(A \cap C)}{P(C)} \\ &= \frac{P(A)}{P(C)} \qquad (\text{since } A \subset C) \\ &= \frac{\frac{1}{4}}{\frac{3}{4}} = \frac{1}{3}. \end{aligned}$$

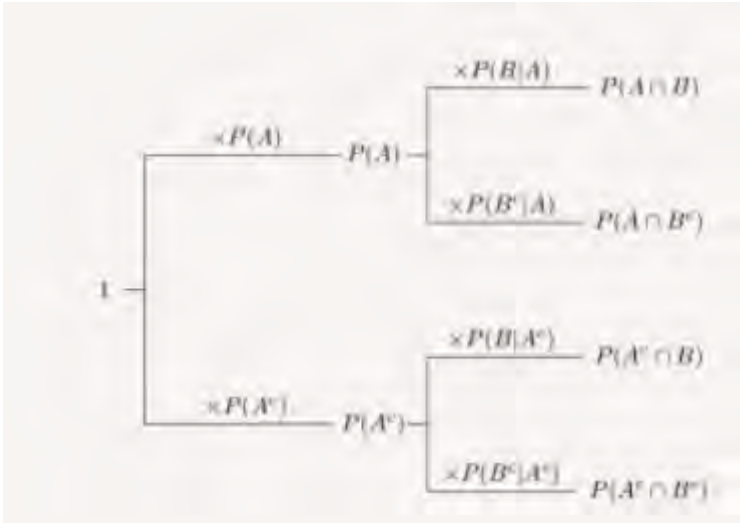
Chain rule for Conditional Probability:

- Proof:

Let us write the formula for conditional probability in the following format

$$P(A \cap B)=P(A)P(B|A)=P(B)P(A|B)$$

This format is particularly useful in situations when we know the conditional probability, but we are interested in the probability of the intersection. We can interpret this formula using a tree diagram such as the one shown below. In this figure, we obtain the probability at each point by multiplying probabilities on the branches leading to that point. This type of diagram can be very useful for some problems.



Now we can extend this formula to three or more events:
$P(A \cap B \cap C)=P(A \cap (B \cap C))=P(A)P(B \cap C|A)$

$P(B \cap C)=P(B)P(C|B)$. Equation (a)
Conditioning both sides on A, we obtain
$P(B \cap C|A)=P(B|A)P(C|A,B)$ Equation (b)

Combining Equation a and b, we obtain the following chain rule:

$$\mathbf{P(A \cap B \cap C)=P(A)P(B|A)P(C|A,B).}$$

Chain rule for conditional probability:

$$\mathbf{P(A_1 \cap A_2 \cap \cdots \cap A_n)=P(A_1)P(A_2|A_1)P(A_3|A_2,A_1)\cdots P(A_n|A_{n-1}A_{n-2}\cdots A_1)}$$

Independence

Two events A and B are independent if and only if P(A∩B)=P(A)P(B).

Now, let's first reconcile this definition with what we mentioned earlier, $P(A|B) = P(A)$. If two events are independent, then $P(A \cap B) = P(A)P(B)$, so

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} \\ &= \frac{P(A)P(B)}{P(B)} \\ &= P(A). \end{aligned}$$

Thus, if two events A and B are independent and $P(B) \neq 0$, then $P(A|B) = P(A)$. To summarize, we can say "independence means we can multiply the probabilities of events to obtain the probability of their intersection", or equivalently, "independence means that conditional probability of one event given another is the same as the original (prior) probability".

- For three events

Three events A, B, and C are independent if **all** of the following conditions hold

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

- In general;

$$P(A_1 \cap A_2 \cap A_3 \cdots \cap A_n) = P(A_1)P(A_2)P(A_3)\cdots P(A_n).$$

- Also We can prove

If $A_1, A_2, \cdots, A_n$ are independent then

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = 1 - (1 - P(A_1))(1 - P(A_2))\cdots(1 - P(A_n)).$$

Difference between Disjointness and Independence.

Concept	Meaning	Formulas
Disjoint	A and B cannot occur at the same time	$A \cap B = \emptyset$, $P(A \cup B) = P(A) + P(B)$
Independent	A does not give any information about B	$P(A\|B) = P(A), P(B\|A) = P(B)$ $P(A \cap B) = P(A)P(B)$

Law of Total Probability

Law of Total Probability:
If $B_1, B_2, B_3, \cdots$ is a partition of the sample space S, then for any event A we have

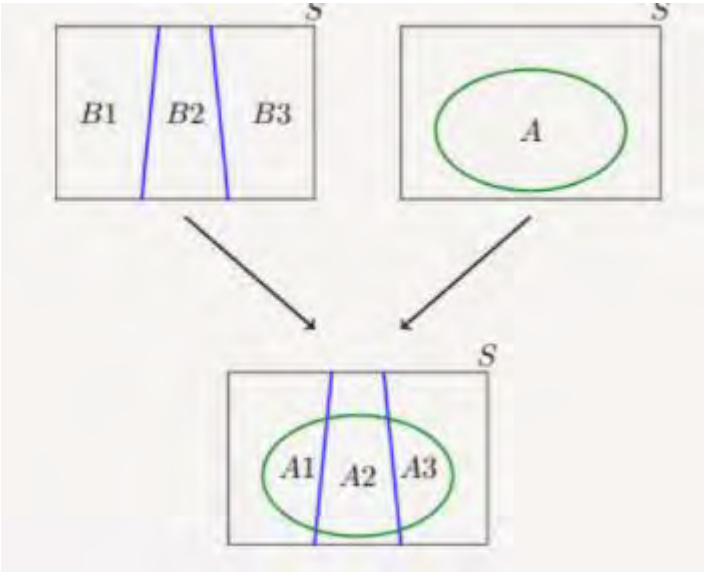
$$P(A) = \sum_i P(A \cap B_i) = \sum_i P(A|B_i)P(B_i).$$

- Using a Venn Diagram, Lets break down the thermo;

$$A_1 = A \cap B_1,$$
$$A_2 = A \cap B_2,$$
$$A_3 = A \cap B_3.$$

- Where;

$$P(A) = P(A_1) + P(A_2) + P(A_3).$$



Here is a proof of the law of total probability using probability axioms:

Proof

Since $B_1, B_2, B_3, \cdots$ is a partition of the sample space S, we can write

$$
\begin{aligned}
S &= \bigcup_i B_i \\
A &= A \cap S \\
&= A \cap \left(\bigcup_i B_i\right) \\
&= \bigcup_i (A \cap B_i) \qquad \text{by the distributive law}
\end{aligned}
$$

Now note that the sets $A \cap B_i$ are disjoint (since the B_i's are disjoint). Thus, by the third probability axiom,

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

Bayes' Rule

- For any two events A and B, where $P(A) \neq 0$, we have

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

- If $B_1, B_2, B_3, \cdots$ form a partition of the sample space S, and A is any event with $P(A) \neq 0$, we have

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_i P(A|B_i)P(B_i)}.$$

Conditional Independence

As we mentioned earlier, almost any concept that is defined for probability can also be extended to conditional probability. Remember that two events A and B are independent if

$$P(A \cap B) = P(A)P(B), \quad \text{or equivalently, } P(A|B) = P(A).$$

We can extend this concept to conditionally independent events. In particular,

Definition
Two events A and B are **conditionally independent** given an event C with $P(C) > 0$ if

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

Recall that from the definition of conditional probability,

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

if $P(B) > 0$. By conditioning on C, we obtain

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

if $P(B|C), P(C) \neq 0$. If A and B are conditionally independent given C, we obtain

$$\begin{aligned} P(A|B, C) &= \frac{P(A \cap B|C)}{P(B|C)} \\ &= \frac{P(A|C)P(B|C)}{P(B|C)} \\ &= P(A|C). \end{aligned}$$

Thus, if A and B are conditionally independent given C, then

$$P(A|B, C) = P(A|C)$$

Lecture 3

Combinatroics

Counting

- For a finite sample space S with equally likely outcomes, the probability of an event A is given by;
- $$P(A)=\frac{|A|}{|S|}=\frac{M}{N}$$
- Thus, finding probability of A reduces to a **counting** problem in which we need to count how many elements are in A and S.
- In this section, we will discuss ways to count the number of elements in a set in an efficient manner.
- Counting is an area of its own and there are books on this subject alone. Here we provide a basic introduction to the material that is usually needed in probability.
- Almost everything that we need about counting is the result of the **multiplication principle**.

Multiplication Principal

Multiplication Principle

Suppose that we perform r experiments such that the kth experiment has n_k possible outcomes, for $k = 1,2,\cdots,r$. 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.

•**Sampling**: sampling from a set means choosing an element from that set. We often **draw** a sample at random from a given set in which each element of the set has equal chance of being chosen.

•**With or without replacement**: usually we draw multiple samples from a set. If we put each object back after each draw, we call this **sampling with replacement**. In this case a single object can be possibly chosen multiple times. For example, if $A=\{a_1,a_2,a_3,a_4\}A=\{a1,a2,a3,a4\}$ and we pick 3 elements with replacement, a possible choice might be (a_3,a_1,a_3). Thus "with replacement" means "repetition is allowed." On the other hand, if repetition is not allowed, we call it **sampling without replacement**.

•**Ordered or unordered**: If ordering matters (i.e.: $a_1,a_2,a_3 \neq a_2,a_3$), this is called **ordered sampling**. Otherwise, it is called **unordered**.

Sampling from sets, we can talk about four possibilities.
•ordered sampling with replacement
•ordered sampling without replacement
•unordered sampling without replacement
•unordered sampling with replacement

Ordered Sampling with Replacement

Here we have a set with n elements (e.g.: $A = \{1, 2, 3, \cdots . n\}$), 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:

1. (1,1);
2. (1,2);
3. (1,3);
4. (2,1);
5. (2,2);
6. (2,3);
7. (3,1);
8. (3,2);
9. (3,3).

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

$$n \times n \times \ldots \times n = n^k$$

Note that this is a special case of the multiplication principle where there are k "experiments" and each experiment has n possible outcomes.

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:

1. (1,2);
2. (1,3);
3. (2,1);
4. (2,3);
5. (3,1);
6. (3,2).

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, ... $(n-k+1)$ options for the kth 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

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

Any of the chosen lists in the above setting (choose k elements, ordered and no repetition) is called a k-permutation of the elements in set A. We use the following notation to show the number of k-permutations of an n-element set:

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

Note that if k is larger than n, then $P_k^n = 0$. This makes sense, since if $k > n$ there is no way to choose k distinct elements from an n-element set. Let's look at a very famous problem, called the birthday problem, or the birthday paradox.

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

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

which is denoted by $n!$, pronounced "n factorial". Thus $n!$ is simply the total number of permutations of n elements, i.e., the total number of ways you can order n different objects. To make our formulas consistent, we define $0! = 1$.

Now, using the definition of $n!$, we can rewrite the formula for P_k^n as

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

The number of k-permutations of n distinguishable objects is given by

$$P_k^n = \frac{n!}{(n-k)!}, \text{ for } 0 \leq k \leq n.$$

Note: There are several different common notations that are used to show the number of k-permutations of an n-element set including $P_{n,k}, P(n,k), nPk$, etc. In this book, we always use P_k^n.

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:

1. {1,2};
2. {1,3};
3. {2,3}.

We show the number of k-element subsets of A by

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

This is read "n choose k." A typical scenario here is that we have a group of n people, and we would like to choose k of them to serve on a committee. 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. In fact, for any k-element subset of $A = \{1, 2, 3, \ldots . n\}$, we can order the elements in $k!$ ways, thus we can write

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

Therefore,

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

Note that if k is an integer larger than n, then $\binom{n}{k} = 0$. This makes sense, since if $k > n$ there is no way to choose k distinct elements from an n-element set.

The number of k-combinations of an n-element set is given by

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}, \text{ for } 0 \leq k \leq n.$$

$\binom{n}{k}$ is also called the **binomial coefficient.** This is because the coefficients in the binomial theorem are given by $\binom{n}{k}$. In particular, the binomial theorem states that for an integer $n \geq 0$, we have

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

Note: There are several different common notations that are used to show the number of k-combinations of an n-element set including $C_{n,k}, C(n,k), C_k^n, nCk$, etc. , we always use $\binom{n}{k}$.

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 "heads" and "tails." You can define "heads" as success and "tails" as "failure" here.
  - •You take a pass-fail test. The possible outcomes are "pass" and "fail."
- 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.

Binomial Formula:
For n independent Bernoulli trials where each trial has success probability p, the probability of k successes is given by

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

Unordered Sampling with Replacement

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

$$x_1 + x_2 + \ldots + x_n = k, \text{ where } x_i \in \{0, 1, 2, 3, \ldots\}.$$

So far we have seen the number of unordered k-samples from an n element set is the same as the number of solutions to the above equation. But how do we find the number of solutions to that equation?

The number of distinct solutions to the equation

$$x_1 + x_2 + \ldots + x_n = k, \text{ where } x_i \in \{0, 1, 2, 3, \ldots\} \qquad (2.3)$$

is equal to

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

SUMMARY

ordered sampling with replacement	n^k
ordered sampling without replacement	$P_k^n = \frac{n!}{(n-k)!}$
unordered sampling without replacement	$\binom{n}{k} = \frac{n!}{k!(n-k)!}$
unordered sampling with replacement	$\binom{n+k-1}{k}$

EXAMPLES

Problem `

Let A and B be two finite sets, with $|A|=m$ and $|B|=n$. How many distinct functions (mappings) can you define from set A to set B, $f:A \rightarrow B$?

Lecture 4
Random Variables

- Definitions, Notation
- Probability Distributions
- Application of Probability Rules
- Mean and s.d. of Random Variables; Rules

Definition

Random Variable: a quantitative variable whose values are results of a random process

Definitions

- **Discrete Random Variable:** one whose possible values are finite or countably infinite (like the numbers 1, 2, 3, …)
- **Continuous Random Variable:** one whose values constitute an entire (infinite) range of possibilities over an interval

Notation

Random Variables are generally denoted with capital letters such as X, Y, or Z.

The letter Z is often reserved for random variables that follow a standardized normal distribution.

Example: *A Simple Random Variable*

- **Background**: Toss a coin twice, and let the random variable X be the number of tails appearing.
- **Questions:**
  - What are the possible values of X?
  - What kind of random variable is X?
- **Responses:**
  - Possible values: 0, 1, 2
  - X is a discrete random variable.

Definitions

- **Probability distribution** of a random variable tells all of its possible values along with their associated probabilities.
- **Probability histogram** displays possible values of a random variable along horizontal axis, probabilities along vertical axis.

Definition

- **Probability distribution** of a random variable tells all of its possible values along with their associated probabilities.

Median and Mean of Probability Distribution

- **Median** is the middle value, with half of values above and half below (equal area value on histogram).
- **Mean** is average value ("balance point" of histogram)
- **Mean equals Median** for symmetric distributions

Example: *Probability Distribution of a Random Variable*

- **Background**: The random variable X is the number of tails in two tosses of a coin.
- **Questions:**
  - What are the probabilities of the possible outcomes?
  - What is the probability distribution of X?
- **Responses:** Possible outcomes:

Each has probability ¼ so the probability distribution is:

X = Number of tails	0	1	2
Probability	1/4	1/2	1/4

Non-overlapping "Or" Rule→P(X=1)=1/2

Example: *Probability Distribution of a Random Variable*

- **Background**: We have the probability distribution of the random variable X for number of tails in two tosses of a coin.

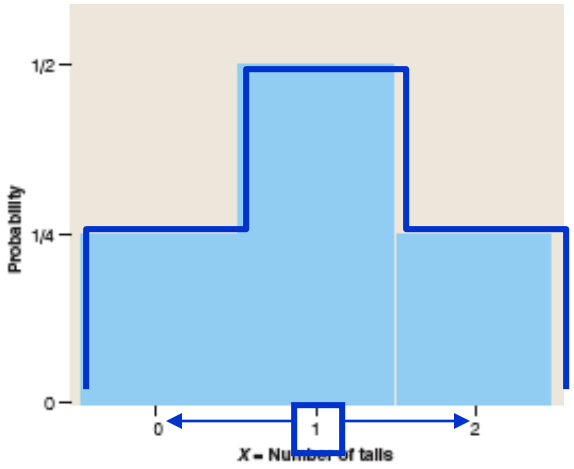
X = Number of tails	0	1	2
Probability	1/4	1/2	1/4

- **Question:** How do we display and summarize X?
- **Response:** Use **probability histogram**.

  Summarize: (center) mean=median=1

  (spread) Typical distance from 1 is a bit less than 1.

  (shape) unimodal, symmetric



Notation; Permissible Probabilities and Sum-to-One Rule for Probability Distributions

P(*X*=*x*) denotes the probability that the random variable X takes the value x.

Any probability distribution of a discrete random variable X must satisfy:

- $0 \leq P(X = x) \leq 1$ where x is any value of X
- $P(X = x_1) + P(X = x_2) + \cdots + P(X = x_k) = 1$ where $x_1, x_2, \cdots, x_k$ are all possible values of X

According to this Rule, if a probability histogram has bars of width 1, their total area must be 1.

Interim Table

To construct probability distribution for more complicated random processes, begin with interim table showing all possible outcomes and their probabilities.

Example: *Interim Table and Probability Distribution*

- **Background:** A coin is tossed 3 times and the random variable X is number of tails tossed.
- **Questions:** What are the possible outcomes, values of X, and probabilities? How do we find probability that $X=1$? $X=2$?
- **Response:**
  - Interim Table:
  - Use Non-overlapping "Or" Rule to combine probabilities

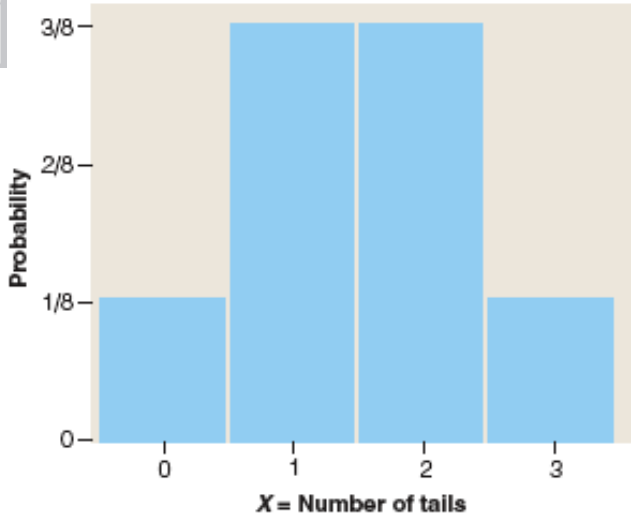
Outcome	X=no.of tails	Probability
HHH	0	1/8
HHT	1	1/8
HTH	1	1/8
THH	1	1/8
HTT	2	1/8
THT	2	1/8
TTH	2	1/8
TTT	3	1/8

Example: *Probability Distribution and Histogram*

- **Background:** X is number of tails in 3 coin tosses.
- **Question:** What are the probability distribution of X and probability histogram?
- **Response:** Use the interim table to determine probabilities.

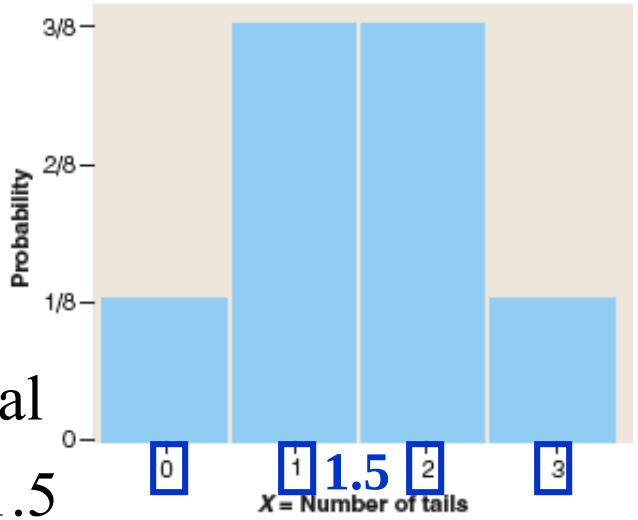
X = Number of tails	0	1	2	3
$P(X = x)$	1/8	3/8	3/8	1/8

Use the probability distribution to sketch the histogram.



Example: *Summaries from Probability Histogram*

- **Background:** Histogram for number of tails in 3 coin tosses.



- **Question:** What does it show?
- **Response:**

  Histogram has

  - Shape: symmetric, unimodal
  - Center: median = mean = 1.5
  - Spread: Typical distance from mean a bit less than 1, since 1 and 2 (which are more common) are only 0.5 away from 1.5; 0 and 3 (less common) are 1.5 away from 1.5.

Definition *(Review)*

- **Probability:** chance of an event occurring, determined as the
  - Proportion of equally likely outcomes comprising the event; or
  - Proportion of outcomes observed in the long run that comprised the event; or
  - Likelihood of occurring, assessed subjectively.

Example: *Different Ways to Assess Probabilities*

- **Background**: Census Bureau reported distribution of U.S. household size in 2000.

X	1	2	3	4	5	6	7
$P(X = x)$	0.26	0.34	0.16	0.14	0.07	0.02	0.01

- **Question:** What is the difference between how these probabilities have been assessed, and the way we assessed probabilities for coin-flip examples?
- **Response:** Coin-flip probabilities are based on known properties of coin (two equally likely faces).

  Household probabilities are based on long-run observed outcomes (all households in U.S. in 2000).

Probability Rules *(Review)*

Probabilities must obey

- Permissible Probabilities Rule
- Sum-to-One Rule
- "Not" Rule
- Non-Overlapping "Or" Rule
- Independent "And" Rule
- General "Or" Rule
- General "And" Rule
- Rule of Conditional Probability

Example: *Permissible Probabilities Rule*

- **Background**: Household size in U.S. has

X	1	2	3	4	5	6	7
$P(X = x)$	0.26	0.34	0.16	0.14	0.07	0.02	0.01

- **Question:** How do these probabilities conform to the Permissible Probabilities Rule?
- **Response:** All between 0 and 1.

Example: *Sum-to-One Rule*

- **Background**: Household size in U.S. has

X	1	2	3	4	5	6	7
$P(X = x)$	0.26	0.34	0.16	0.14	0.07	0.02	0.01

- **Question:** According to the "Sum-to-One" Rule, what must be true about the probabilities in the distribution?
- **Response:** According to the Rule, we have 0.26+0.34+0.16+0.14+0.07+0.02+0.01=1

Example: *"Not" Rule*

- **Background**: Household size in U.S. has

X	1	2	3	4	5	6	7
$P(X = x)$	0.26	0.34	0.16	0.14	0.07	0.02	0.01

- **Question:** According to the "Not" Rule, what is the probability of a household *not* consisting of just one person?
- **Response:**

$$P(X \neq 1) = 1 - P(X = 1) = 1 - 0.26 = 0.74$$

Example: *Non-Overlapping "Or" Rule*

- **Background**: Household size in U.S. has

X	1	2	3	4	5	6	7
$P(X = x)$	0.26	0.34	0.16	0.14	0.07	0.02	0.01

- **Question:** According to the Non-overlapping "Or" Rule, what is the probability of having fewer than 3 people?
- **Response:** The probability of having fewer than 3 people is $P(X<3)$

  $=P(X=1 \text{ or } X=2)=P(X=1)+P(X=2)=0.26+0.34=0.60$

Example: *Independent "And" Rule*

- **Background**: Household size in U.S. has

X	1	2	3	4	5	6	7
$P(X = x)$	0.26	0.34	0.16	0.14	0.07	0.02	0.01

- **Question:** Suppose a polling organization has sampled two households at random. According to the Independent "And" Rule, what is the probability that the first has 3 people and the second has 4 people?
- **Response:** The probability that the first has 3 people and the second has 4 people is

P($X1$=3 and $X2$=4)

=P($X1$=3)×P($X2$=4)=(0.16)(0.14)=0.0224

where we use $X1$ to denote number in 1st household,
$X2$ to denote number in 2nd household.

Example: *General "Or" Rule*

- **Background**: Household size in U.S. has

X	1	2	3	4	5	6	7
$P(X = x)$	0.26	0.34	0.16	0.14	0.07	0.02	0.01

- **Question:** Suppose a polling organization has sampled two households at random. According to the General "Or" Rule, what is the probability that one or the other has 3 people?
- **Response:** The events overlap: it is possible that both households have 3 people. P(*X*1=3 or *X*2=3) =

P(*X*1=3)+P(*X*2=3)-P(*X*1=3 and *X*2=3) =
0.16+0.16-(0.16)(0.16) = 0.2944

where we apply the Independent "And" Rule for P(*X*1=3 and *X*2=3).

Example: *Rule of Conditional Probability*

- **Background**: Household size in U.S. has

X	1	2	3	4	5	6	7
$P(X = x)$	0.26	0.34	0.16	0.14	0.07	0.02	0.01

- **Question:** Suppose a polling organization samples only from households with fewer than 3 people. What is the probability that a household with fewer than 3 people has only 1 person?
- **Response:** P(X=1 given X<3) =

$$\frac{\text{P}(X=1 \text{ and } X<3)}{\text{P}(X<3)} = \frac{0.26}{0.26+0.34} = 0.43$$

Mean and Standard Deviation of Random Variable

- **Mean of discrete random variable *X***

$$\mu = x_1 P(X = x_1) + \cdots + x_k P(X = x_k)$$

Mean is **weighted average** of values, where each value is weighted with its probability.

- **Standard deviation of discrete random variable *X***

$$\sigma = \sqrt{(x_1 - \mu)^2 P(X = x_1) + \cdots + (x_k - \mu)^2 P(X = x_k)}$$

Standard deviation is "typical" distance of values from mean. Squared standard deviation is the **variance**.

Looking Back: *Greek letters are used because these are the mean and standard deviation of* ***all*** *the random variables' values.*

Example: *Mean of Random Variable*

- **Background**: Household size in U.S. has

X	1	2	3	4	5	6	7
$P(X = x)$	0.26	0.34	0.16	0.14	0.07	0.02	0.01

- **Question:** What is the mean household size?
- **Response:** $1(0.26)+2(0.34)+\ldots 7(0.01) = 2.5$ is the mean household size.

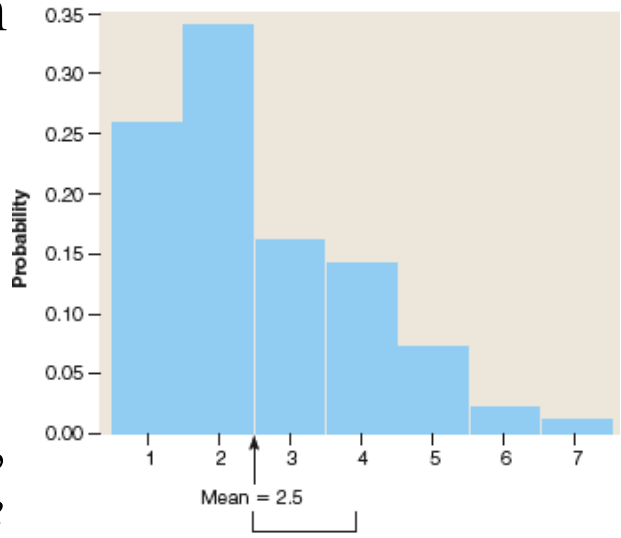
Looking Back: *Median is 2 (has 0.5 at or below it). Mean is greater than median because distribution is skewed right. Also, mean is less than the "middle" number, 4, because smaller household sizes are weighted with higher probabilities.*

Example: *Standard Deviation of R.V.*

- **Background**: Household size in U.S. has

X	1	2	3	4	5	6	7
$P(X = x)$	0.26	0.34	0.16	0.14	0.07	0.02	0.01

- **Question:** What is the standard deviation of household sizes (typical distance from the mean, 2.5)?
  (a) 0.014 (b) 0.14 (c) 1.4 (d) 14.0
- **Response:** The typical distance of household sizes from their mean, 2.5, is 1.4: the closest are 0.5 away (2 and 3), the farthest is 4.5 away (7). *(Or calculate by hand or with software).*



Standard deviation =1.4

A Closer Look: *Skewed right → most of the spread arises from values above the mean, not below.*

Rules for Mean and Variance

- Multiply R.V. by constant→its mean and standard deviation are multiplied by same constant [or its abs. value, since s.d.>0]
- Take sum of two independent R.V.s→
  - mean of sum = sum of means
  - variance of sum = sum of variances

(variance is *squared* standard deviation)

Looking Ahead: *These rules will help us identify mean and standard deviation of sample proportion and sample mean.*

Example: *Mean, Variance, and SD of R.V.*

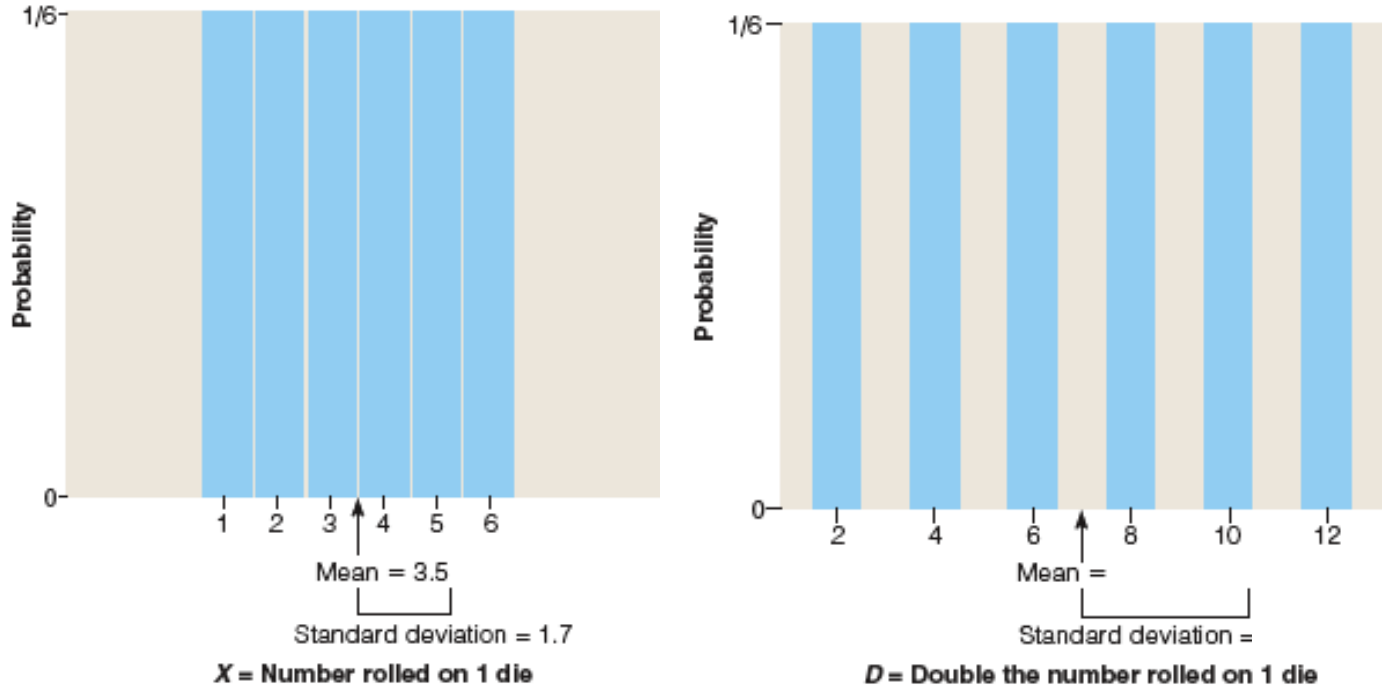
- **Background**: Number *X* rolled on a die has

X=no. rolled	1	2	3	4	5	6
P(X=x)	1/6	1/6	1/6	1/6	1/6	1/6

- **Question:** What are the mean, variance, and standard deviation of *X*?
- **Response:**
  - **Mean:** same as median 3.5 (because symmetric)
  - **Variance:** 2.92 (found by hand or with software)
  - **Standard deviation:** 1.7 (square root of variance)

Example: *Mean and SD for Multiple of R.V.*

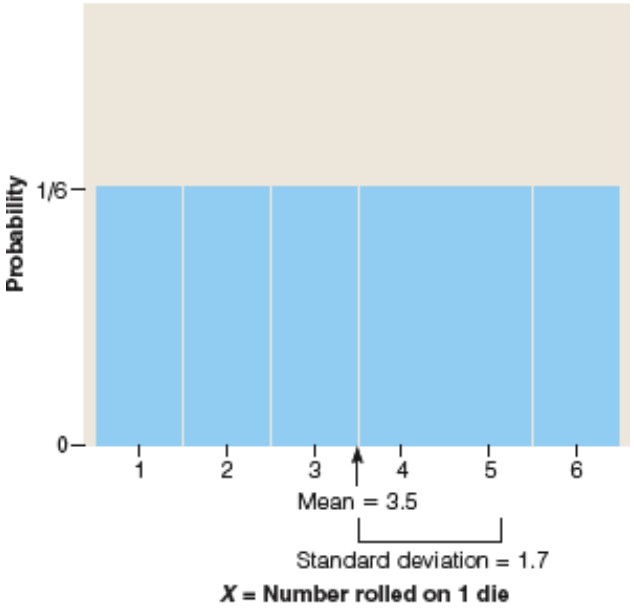
- **Background**: Number *X* rolled on a die has mean 3.5, s.d. 1.7.

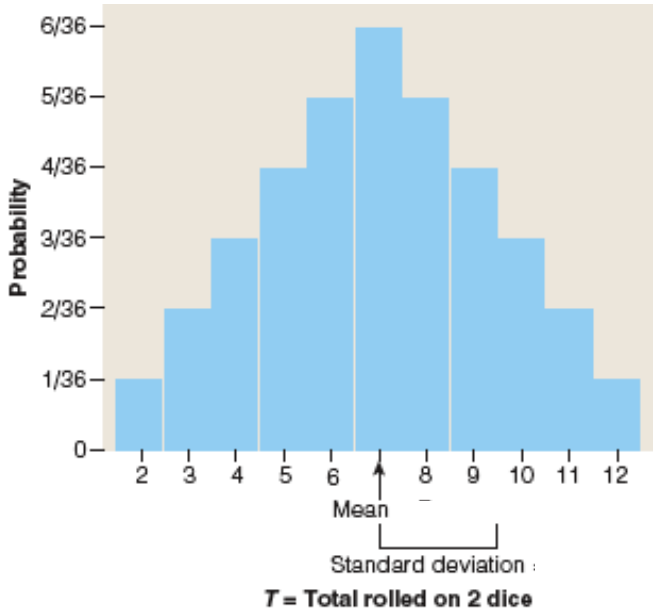


- **Question:** What are mean and s.d. of double the roll?
- **Response:** For double the roll, mean is $2(3.5) = 7$, s.d. is $2(1.7) = 3.4$.

Example: *Mean and SD for Sum of R.V.s*

- **Background**: Numbers $X1$, $X2$ on 2 dice each have mean 3.5, variance 2.92.

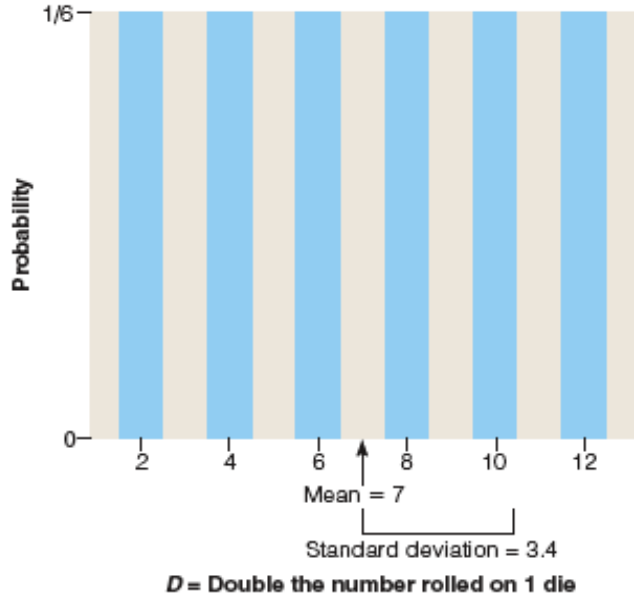


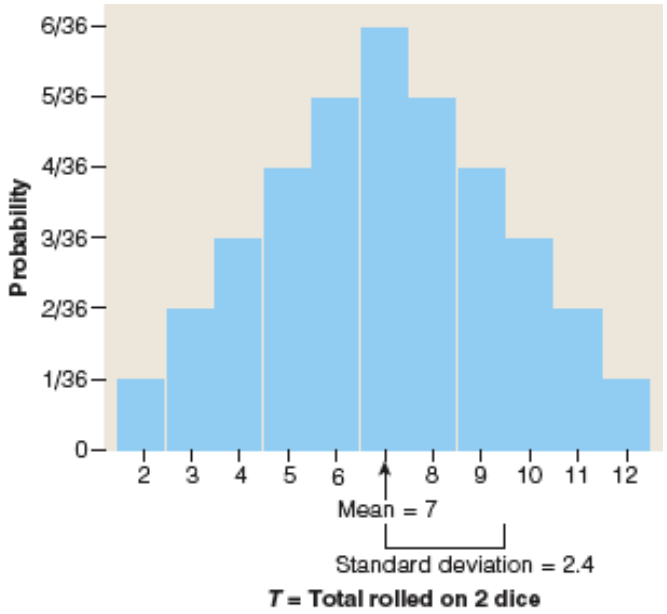


- **Question:** What are mean, variance, and s.d. of total on 2 dice?
- **Response**: Mean $3.5 + 3.5 = 7$, variance $2.92 + 2.92 = 5.84$, s.d. square root of $5.84 = 2.4$.

Example: *Doubling R.V. or Adding Two R.V.s*

- **Background**: Double roll of a die: mean=7, s.d.= 3.4.
  Total of 2 dice: mean=7, s.d.= 2.4.

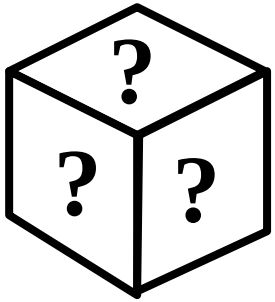




- **Question**: Why is double roll more spread than total of 2 dice?
- **Response**: Doubling roll of 1 die makes extremes [2(1)=2 or 2(6)=12] more likely; totaling 2 dice tends to have low and high rolls "cancel each other out".

Example: *Doubling R.V. or Adding Two R.V.s*

- This is the key to the benefits of sampling many individuals: The average of their responses gets us closer to what's true for the larger group.
- If the numbers on a die were unknown, and you had to guess their mean value, would you make a better guess with a single roll or the average of two rolls?



Lecture Summary
(Random Variables)

- Random variables
  - Discrete vs. continuous
  - Notation
- Probability distributions: displaying, summarizing
- Probability rules applied to random variables
- Constructing distribution table
- Mean and standard deviation of random variable
- Rules for mean and variance

Discrete Random Variables

Discrete Random Variables

- There are two important classes of random variables that we discussed earlier: *discrete random variables* and *continuous random variables*.
- We will discuss discrete random variables in this lecture.
- There will be a third class of random variables that are called *mixed random variables*.
- Mixed random variables, as the name suggests, can be thought of as mixture of discrete and continuous random variables. We will discuss mixed random variables in lectures.

- Remember that a set A is countable if either
  - A is a finite set such as {1,2,3,4}, or
  - It can be put in one-to-one correspondence with natural numbers (in this case the set is said to be countably infinite)
  - In particular, as we discussed in previous lectures, sets such as N,Z,Q and their subsets are countable, while sets such as nonempty intervals [a,b] in R are uncountable.

- A random variable is discrete if its range is a countable set.

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 RX. In other words, we can write
$R_x=\{x1,x2,x3,..\}$.

Note that here $x_1,x_2,x_3,...$ 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 such as x, x_1, y, z, etc.

For a discrete random variable X, we are interested in knowing the probabilities of $X=x_k$. Note that here, the event $A=\{X=x_k\}$ is defined as the set of outcomes s in the sample space S for which the corresponding value of X is equal to x_k.

In particular,

$$A=\{s\in S|X(s)=x_k\}.$$

The probabilities of events $\{X=x_k\}\{X=x_k\}$ are formally shown by the **probability mass function (pmf)** of X.

Definition

- Let X be a discrete random variable with range $R_X=\{x1,x2,x3,...\}$ (finite or countably infinite).

  The function

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

is called the *probability mass function (PMF)* of X.

Thus, the PMF is a probability measure that gives us probabilities of the possible values for a random variable. While the above notation is the standard notation for the PMF of X, it might look confusing at first. The subscript X here indicates that this is the PMF of the random variable X. Thus, for example, Px(1) shows the probability that X=1.

Example

I toss a fair coin twice, and let X be defined as the number of heads I observe. Find the range of X, R_X, as well as its probability mass function P_X.

Here, our sample space is given by

$$S = \{HH, HT, TH, TT\}.$$

The number of heads will be 0, 1 or 2. Thus

$$R_X = \{0, 1, 2\}.$$

Since this is a finite (and thus a countable) set, the random variable X is a discrete random variable. Next, we need to find PMF of X. The PMF is defined as

$$P_X(k) = P(X = k) \text{ for } k = 0, 1, 2.$$

We have

$$P_X(0) = P(X = 0) = P(TT) = \frac{1}{4},$$

$$P_X(1) = P(X = 1) = P(\{HT, TH\}) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2},$$

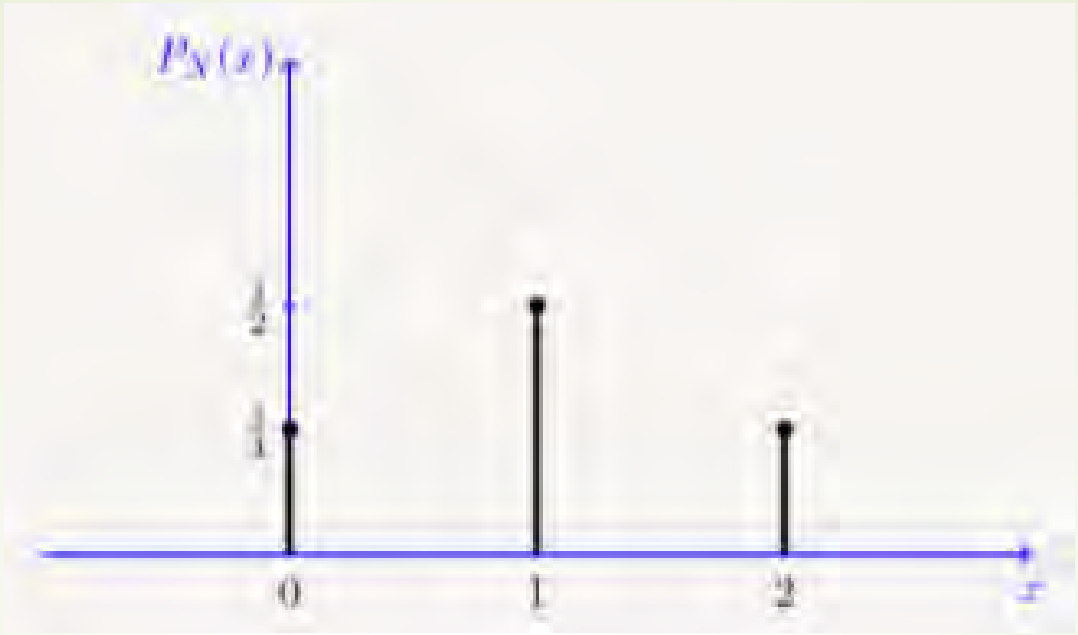
$$P_X(2) = P(X = 2) = P(HH) = \frac{1}{4}.$$

Furthermore;

- Although the PMF is usually defined for values in the range, it is sometimes convenient to extend the PMF of XX to all real numbers. If $x \notin R_X$, we can simply write $P_X(x)=P(X=x)=0$. Thus, in general we can write

$$P_X(x) = \begin{cases} P(X=x) & \text{if } x \text{ is in } R_X \\ 0 & \text{otherwise} \end{cases}$$

To better visualize the PMF, we can plot it. Look at the next slide, where the PMF of the above random variable X. As we see, the random variable can take three possible values 0, 1 and 2. The figure also clearly indicates that the event X=1 is twice as likely as the other two possible values. The Figure can be interpreted in the following way: If we repeat the random experiment (tossing a coin twice) a large number of times, then about half of the times we observe X=1, about a quarter of times we observe X=0, and about a quarter of times we observe X=2.



For discrete random variables, the PMF is also called the **probability distribution**. Thus, when asked to find the probability distribution of a discrete random variable X, we can do this by finding its PMF.

The phrase *distribution function* is usually reserved exclusively for the cumulative distribution function CDF (as defined later in the book). The word *distribution*, on the other hand, in this book is used in a broader sense and could refer to PMF, probability density function (PDF), or CDF.

Example

I have an unfair coin for which P(H)=p, where 0<p<1. I toss the coin repeatedly until I observe a heads for the first time. Let Y be the total number of coin tosses. Find the distribution of Y.

First, we note that the random variable Y can potentially take any positive integer, so we hav $R_Y = \mathbb{N} = \{1, 2, 3, \dots\}$. To find the distribution of Y, we need to find $P_Y(k) = P(Y = k)$ fo $k = 1, 2, 3, \dots$. We have

$$P_Y(1) = P(Y = 1) = P(H) = p,$$

$$P_Y(2) = P(Y = 2) = P(TH) = (1-p)p,$$

$$P_Y(3) = P(Y = 3) = P(TTH) = (1-p)^2 p,$$

$$\vdots \quad \vdots \quad \vdots \quad \vdots$$

$$P_Y(k) = P(Y = k) = P(TT\dots TH) = (1-p)^{k-1} p.$$

Thus, we can write the PMF of Y in the following way

$$P_Y(y) = \begin{cases} (1-p)^{y-1} p & \text{for } y = 1, 2, 3, \dots \\ 0 & \text{otherwise} \end{cases}$$

Note

- Consider a discrete random variable X with Range(X)=R_X. Note that by definition the PMF is a probability measure, so it satisfies all properties of a probability measure. In particular, we have
  - $0 \leq P_X(x) \leq 1$ for all x, and
  - $\sum_{x \in R_X} P_X(x) = 1$.
- Also note that for any set $A \subset R_X$, we can find the probability that $X \in A$ using the PMF

$$P(X \in A) = \sum_{x \in A} P_X(x).$$

Properties of PMF:

- $0 \leq P_X(x) \leq 1$ for all x;
- $\sum_{x \in R_X} P_X(x) = 1$;
- for any set $A \subset R_X$, $P(X \in A) = \sum_{x \in A} P_X(x)$.

Example

For the random variable Y in previous example,

1. Check that $\sum_{y\in R} YP_Y(y)=1$.
2. If p=1/2, find $P(2\leq Y<5)$.

$$P_Y(k) = P(Y = k) = (1-p)^{k-1}p, \text{ for } k = 1, 2, 3, \ldots$$

Thus,

1. to check that $\sum_{y\in R_Y} P_Y(y) = 1$, we have

$$\begin{aligned}\sum_{y\in R_Y} P_Y(y) &= \sum_{k=1}^{\infty}(1-p)^{k-1}p \\ &= p\sum_{j=0}^{\infty}(1-p)^{j} \\ &= p\frac{1}{1-(1-p)} \qquad \text{Geometric sum} \\ &= 1;\end{aligned}$$

2. if $p = \frac{1}{2}$, to find $P(2 \leq Y < 5)$, we can write

$$\begin{aligned}P(2 \leq Y < 5) &= \sum_{k=2}^{4} P_Y(k) \\ &= \sum_{k=2}^{4}(1-p)^{k-1}p \\ &= \frac{1}{2}\left(\frac{1}{2}+\frac{1}{4}+\frac{1}{8}\right) \\ &= \frac{7}{16}.\end{aligned}$$

Independent Random Variables

- In real life, we usually need to deal with more than one random variable. For example, if you study physical characteristics of people in a certain area, you might pick a person at random and then look at his/her weight, height, etc. The weight of the randomly chosen person is one random variable, while his/her height is another one. Not only do we need to study each random variable separately, but also we need to consider if there is *dependence* (i.e., correlation) between them. Is it true that a taller person is more likely to be heavier or not? The issues of dependence between several random variables will be studied in detail later on, but here we would like to talk about a special scenario where two random variables are *independent*.
- The concept of independent random variables is very similar to independent events. Remember, two events A and B are independent if we have $P(A,B)=P(A)P(B)$ (remember comma means *and*, i.e., $P(A,B)=P(A \text{ and } B)=P(A \cap B)$. Similarly, we have the following definition for independent discrete random variables.

Definition

Consider two discrete random variables X and Y. We say that X and Y are independent if

$$P\Big(X = x, Y = y\Big) = P(X = x)P(Y = y), \qquad \text{for all } x, y.$$

In general, if two random variables are independent, then you can write

$$P\Big(X \in A, Y \in B\Big) = P(X \in A)P(Y \in B), \qquad \text{for all sets } A \text{ and } B.$$

Intuitively, two random variables X and Y are independent if knowing the value of one of them does not change the probabilities for the other one. In other words, if X and Ym are independent, we can write

$$P(Y=y|X=x)=P(Y=y), \text{ for all } x,y..$$

Similar to independent events, it is sometimes easy to argue that two random variables are independent simply because they do not have any physical interactions with each other. Here is a simple example: I toss a coin 2N times. Let X be the number of heads that I observe in the first N coin tosses and let Y be the number of heads that I observe in the second N coin tosses. Since X and Y are the result of independent coin tosses, the two random variables X and Y are independent. On the other hand, in other scenarios, it might be more complicated to show whether two random variables are independent.

Example

I toss a coin twice and define X to be the number of heads I observe. Then, I toss the coin two more times and define Y to be the number of heads that I observe this time. Find P((X<2) and (Y>1)).

Since X and Y are the result of different independent coin tosses, the two random variables X and Y are independent

$$
\begin{aligned}
P\Big((X<2) \text{ and } (Y>1)\Big) &= P(X<2)P(Y>1) \qquad \text{(because } X \text{ and } Y \text{ are independent)} \\
&= \big(P_X(0)+P_X(1)\big)P_Y(2) \\
&= \left(\tfrac{1}{4}+\tfrac{1}{2}\right)\tfrac{1}{4} \\
&= \tfrac{3}{16}.
\end{aligned}
$$

We can extend the definition of independence to n random variables

Consider n discrete random variables $X_1, X_2, X_3, \ldots, X_n$. We say that $X_1, X_2, X_3, \ldots, X_n$ are independent if

$$P\Big(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\Big)$$

$$= P(X_1 = x_1)P(X_2 = x_2)\ldots P(X_n = x_n), \qquad \text{for all } x_1, x_2, \ldots, x_n.$$

Special Distribution

- As it turns out, there are some specific distributions that are used over and over in practice, thus they have been given special names. There is a random experiment behind each of these distributions. Since these random experiments model a lot of real life phenomenon, these special distributions are used frequently in different applications. That's why they have been given a name and we devote a section to study them. We will provide PMFs for all of these special random variables, but rather than trying to memorize the PMF, you should understand the random experiment behind each of them. If you understand the random experiments, you can simply derive the PMFs when you need them. Although it might seem that there are a lot of formulas in this section, there are in fact very few new concepts. Do not get intimidated by the large number of formulas, look at each distribution as a practice problem on discrete random variables.

Bernoulli Distribution

- What is the simplest discrete random variable (i.e., simplest PMF) that you can imagine? My answer to this question is a PMF that is nonzero at only one point. For example, if you define:

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

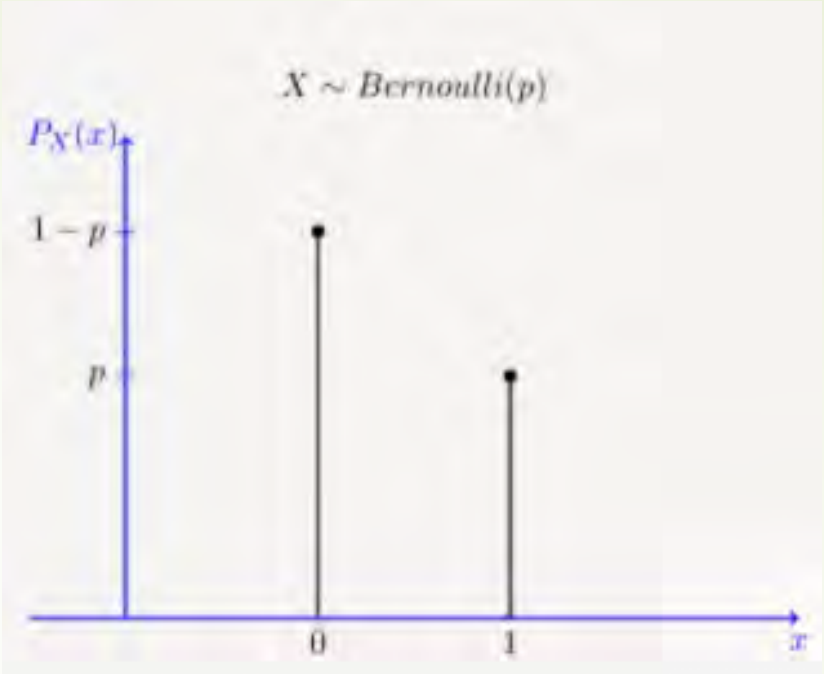
- Then X is a discrete random variable that can only take one value, i.e., X=1 with a probability of one. But this is not a very interesting distribution because it is not actually random. Then, you might ask what is the next simplest discrete distribution. And my answer to that is the **Bernoulli** distribution. A Bernoulli random variable is a random variable that can only take two possible values, usually 0 and 1. This random variable models random experiments that have two possible outcomes, sometimes referred to as "success" and "failure." Here are some examples:
    - •You take a pass-fail exam. You either pass (resulting in X=1) or fail (resulting in X=0).
    - •You toss a coin. The outcome is either heads or tails.
    - •A child is born. The gender is either male or female.

Formally, the Bernoulli distribution is defined as follows:

A random variable X is said to be a *Bernoulli* random variable with *parameter* p, shown as $X \sim Bernoulli(p)$, if its PMF is given by

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

where $0 < p < 1$.



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. In particular, the indicator random variable I_A for an event A is defined by

$$I_A = \begin{cases} 1 & \text{if the event } A \text{ occurs} \\ 0 & \text{otherwise} \end{cases}$$

The indicator random variable for an event A has Bernoulli distribution with parameter $p = P(A)$, so we can write

$$I_A \sim Bernoulli\big(P(A)\big).$$

The random experiment behind the geometric distribution is as follows. 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.

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

We usually define q=1−p, so we can write $P_{X(k)=pq}$ k−1, for k=1,2,3,... To say that a random variable has geometric distribution with parameter p, we write X~Geometric(p).

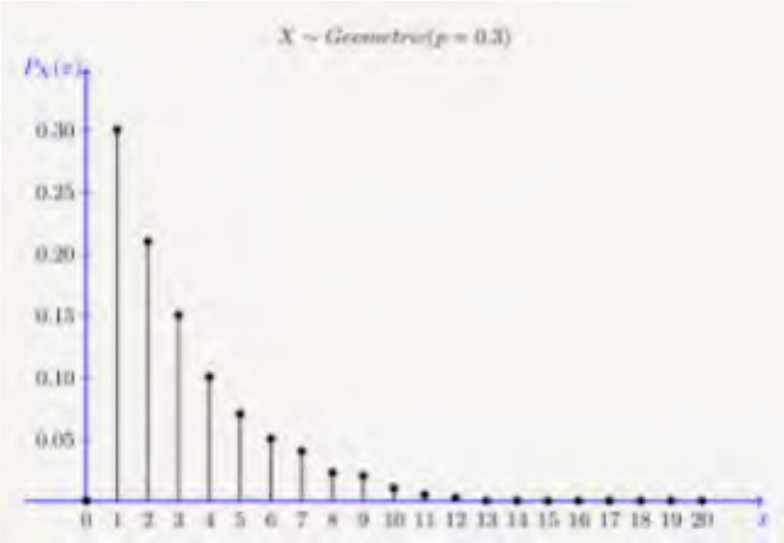
More formally, we have the following definition:

A random variable X is said to be a *geometric* random variable with *parameter* p, shown as $X \sim Geometric(p)$, if its PMF is given by

$$P_X(k) = \begin{cases} p(1-p)^{k-1} & \text{for } k = 1, 2, 3, \ldots \\ 0 & \text{otherwise} \end{cases}$$

where $0 < p < 1$.

Figure shows the PMF of a Geometric(0.3) random variable.



We should note that some books define geometric random variables slightly differently. They define the geometric random variable X as the total number of failures before observing the first success. By this definition the range of X is $R_{X=\{0,1,2,...\}}$ and the PMF is given by

$$P_X(k) = \begin{cases} p(1-p)^k & \text{for } k = 0, 1, 2, 3, \dots \\ 0 & \text{otherwise} \end{cases}$$

In this book, whenever we write X~Geometric(p), we always mean X as the total number of trials. Note that as long as you are consistent in your analysis, it does not matter which definition you use. That is why we emphasize that you should understand how to derive PMFs for these random variables rather than memorizing them.

Binomial Distribution

- The random experiment behind the binomial distribution is as follows. 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, and we write X~Binomial(n,p). The range of X in this case is R_X={0,1,2,...,n}. The PMF of X in this case is given by binomial formula

$$P_X(k) = \binom{n}{k} p^k (1-p)^{n-k}, \text{ for } k = 0, 1, 2, \ldots, n.$$

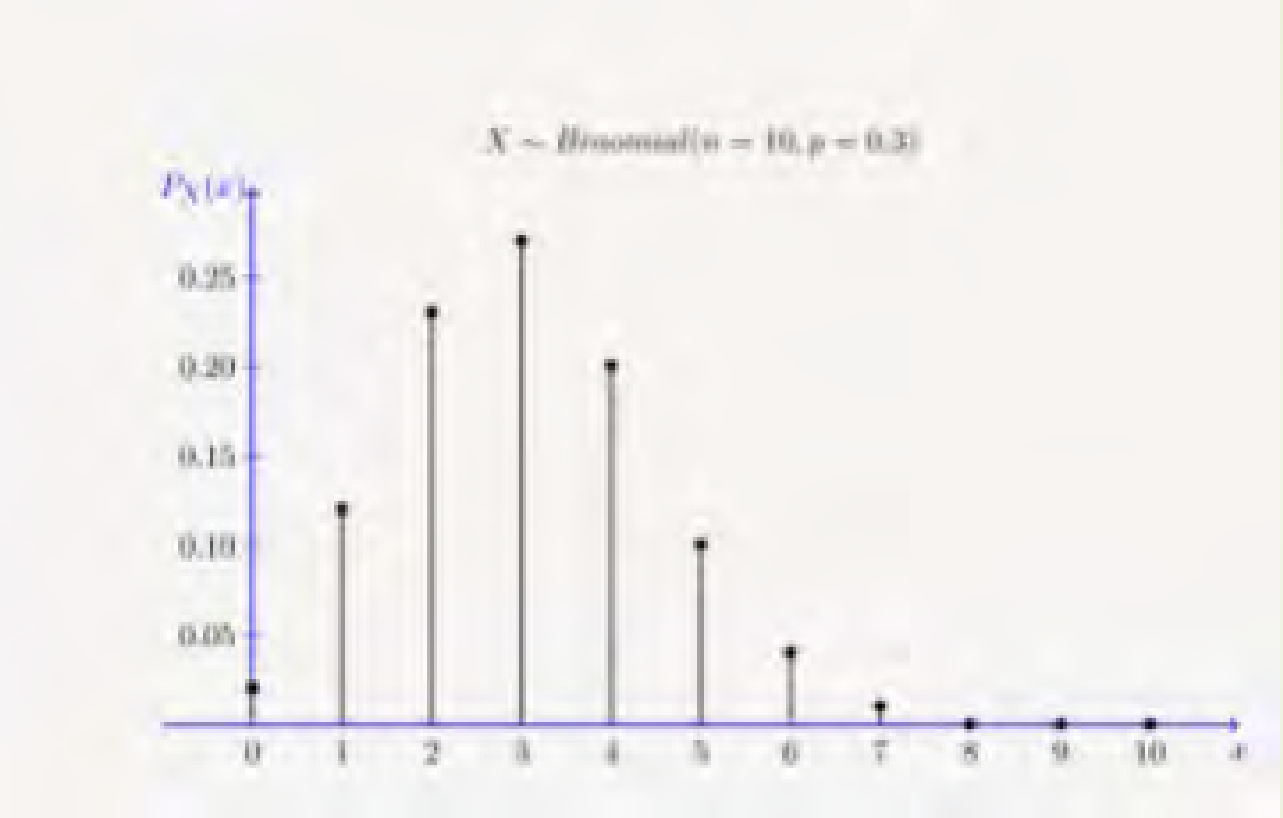
- Definition

A random variable X is said to be a *binomial* random variable with parameters n and p, shown as $X \sim Binomial(n, p)$, if its PMF is given by

$$P_X(k) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & \text{for } k = 0, 1, 2, \cdots, n \\ 0 & \text{otherwise} \end{cases}$$

where $0 < p < 1$.

Binomial(n,p) PMF for n=10, p=0.3 and n=20, p=0.6 respectively.



Binomial random variable as a sum of Bernoulli random variables

Here is a useful way of thinking about a binomial random variable. Note that a Binomial(n,p) random variable can be obtained by n independent coin tosses. If we think of each coin toss as a Bernoulli(p) 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.

Lemma

If X1,X2,...,Xn are independent Bernoulli(p) random variables, then the random variable X defined by X=X1+X2+...+Xn has a Binomial(n,p) distribution.

To generate a random variable X~Binomial(n,p), we can toss a coin n times and count the number of heads. Counting the number of heads is exactly the same as finding X1+X2+...+Xn, where each Xi is equal to one if the corresponding coin toss results in heads and zero otherwise. This interpretation of binomial random variables is sometimes very helpful. Let's look at an example.

- 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.

Since $X \sim Binomial(n,p)$, we can think of X as the number of heads in n independent coin tosses, i.e., we can write

$$X = X_1 + X_2 + \ldots + X_n,$$

where the X_i's are independent $Bernoulli(p)$ random variables. Similarly, since $Y \sim Binomial(m,p)$, we can think of Y as the number of heads in m independent coin tosses, i.e., we can write

$$Y = Y_1 + Y_2 + \ldots + Y_m,$$

where the Y_j's are independent $Bernoulli(p)$ random variables. Thus, the random variable $Z = X + Y$ will be the total number of heads in $n+m$ independent coin tosses:

$$Z = X + Y = X_1 + X_2 + \ldots + X_n + Y_1 + Y_2 + \ldots + Y_m,$$

where the X_i's and Y_j's are independent $Bernoulli(p)$ random variables. Thus, by Lemma 3.1, Z is a binomial random variable with parameters $m+n$ and p, i.e., $Binomial(m+n,p)$. Therefore, the PMF of Z is

$$P_Z(k) = \begin{cases} \binom{m+n}{k} p^k (1-p)^{m+n-k} & \text{for } k = 0,1,2,3,\ldots,m+n \\ 0 & \text{otherwise} \end{cases}$$

where the X_i's and Y_j's are independent $Bernoulli(p)$ random variables. Thus, by Lemma , Z is a binomial random variable with parameters $m+n$ and p, i.e., $Binomial(m+n,p)$. Therefore, the PMF of Z is

$$P_Z(k) = \begin{cases} \binom{m+n}{k} p^k (1-p)^{m+n-k} & \text{for } k = 0,1,2,3,\ldots,m+n \\ 0 & \text{otherwise} \end{cases}$$

The above solution is elegant and simple, but we may also want to directly obtain the PMF of Z using probability rules. Here is another method to solve Example First, we note that $R_Z = \{0,1,2,\ldots,m+n\}$. For $k \in R_Z$, we can write

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

We will find $P(X+Y=k)$ by using conditioning and the law of total probability. In particular, we can write

$$\begin{aligned} P_Z(k) &= P(X+Y=k) \\ &= \sum_{i=0}^{n} P(X+Y=k|X=i)P(X=i) && \text{(law of total probability)} \\ &= \sum_{i=0}^{n} P(Y=k-i|X=i)P(X=i) \\ &= \sum_{i=0}^{n} P(Y=k-i)P(X=i) && \text{(since } X \text{ and } Y \text{ are independent)} \\ &= \sum_{i=0}^{n} \binom{m}{k-i} p^{k-i}(1-p)^{m-k+i} \binom{n}{i} p^i (1-p)^{n-i} && \text{(since } X \text{ and } Y \text{ are binomial)} \\ &= \sum_{i=0}^{n} \binom{m}{k-i} \binom{n}{i} p^k (1-p)^{m+n-k} \\ &= p^k (1-p)^{m+n-k} \sum_{i=0}^{n} \binom{m}{k-i} \binom{n}{i} \\ &= \binom{m+n}{k} p^k (1-p)^{m+n-k} \end{aligned}$$

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 mm successes. Again, different authors define the Pascal distribution slightly differently, and as we mentioned before if you understand one of them you can easily derive the other ones. Here is how we define the Pascal distribution in this book. Suppose that I have a coin with P(H)=p. I toss the coin until I observe mm heads, where $m \in N$. 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). Note that by our definition the range of X is given by $R_X=\{m, m+1, m+2, m+3, \cdots\}$.
- Let us derive the PMF of a Pascal(m,p) random variable X. Suppose that I toss the coin until I observe mm heads, and X is defined as the total number of coin tosses in this experiment.

- To find the probability of the event A={X=k}, we argue as follows. By definition, event A can be written as A=B∩C, where
  - B is the event that we observe m−1 heads (successes) in the first k−1 trials, and
  - C is the event that we observe a heads in the kth trial.
- Note that B and C are independent events because they are related to different independent trials (coin tosses). Thus we can write
- P(A)=P(B∩C)=P(B)P(C).
- Now, we have P(C)=p. Note also that P(B) is the probability that I observe m− heads in the k−1 coin tosses. This probability is given by the binomial formula, in particular

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

Thus, we obtain

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

A random variable X is said to be a *Pascal* random variable with parameters m and p, shown as $X \sim Pascal(m, p)$, if its PMF is given by

$$P_X(k) = \begin{cases} \binom{k-1}{m-1} p^m (1-p)^{k-m} & \text{for } k = m, m+1, m+2, m+3, \ldots \\ 0 & \text{otherwise} \end{cases}$$

where $0 < p < 1$.

Hypergeometric Distribution

Here is the random experiment behind the hypergeometric 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 the 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 have

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

A random variable X is said to be a *Hypergeometric* random variable with parameters b, r and k, shown as $X \sim Hypergeometric(b, r, k)$, if its range is $R_X = \{\max(0, k-r), \max(0, k-r)+1, \max(0, k-r)+2, \ldots, \min(k, b)\}$, and its PMF is given by

$$P_X(x) = \begin{cases} \frac{\binom{b}{x}\binom{r}{k-x}}{\binom{b+r}{k}} & \text{for } x \in R_X \\ 0 & \text{otherwise} \end{cases}$$

Again, there is no point to memorizing the PMF. All you need to know is how to solve problems that can be formulated as a hypergeometric random variable.

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.

1. What is the probability that I get no emails in an interval of length 5 minutes?
2. What is the probability that I get more than 3 emails in an interval of length 10 minutes?

1. Let X be the number of emails that I get in the 5-minute interval. Then, by the assumption X is a Poisson random variable with parameter $\lambda = 5(0.2) = 1$,

$$P(X = 0) = P_X(0) = \frac{e^{-\lambda}\lambda^0}{0!} = \frac{e^{-1} \cdot 1}{1} = \frac{1}{e} \approx 0.3679$$

2. Let Y be the number of emails that I get in the 10-minute interval. Then by the assumption Y is a Poisson random variable with parameter $\lambda = 10(0.2) = 2$,

$$\begin{aligned} P(Y > 3) &= 1 - P(Y \leq 3) \\ &= 1 - \left(P_Y(0) + P_Y(1) + P_Y(2) + P_Y(3)\right) \\ &= 1 - e^{-\lambda} - \frac{e^{-\lambda}\lambda}{1!} - \frac{e^{-\lambda}\lambda^2}{2!} - \frac{e^{-\lambda}\lambda^3}{3!} \\ &= 1 - e^{-2} - \frac{2e^{-2}}{1} - \frac{4e^{-2}}{2} - \frac{8e^{-2}}{6} \\ &= 1 - e^{-2}\left(1 + 2 + 2 + \frac{8}{6}\right) \\ &= 1 - \frac{19}{3e^2} \approx 0.1429 \end{aligned}$$

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 n is very large and 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.

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

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

Proof

We have

$$\begin{aligned}
\lim_{n\to\infty} P_X(k) &= \lim_{n\to\infty} \binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k} \\
&= \lambda^k \lim_{n\to\infty} \frac{n!}{k!(n-k)!}\left(\frac{1}{n^k}\right)\left(1-\frac{\lambda}{n}\right)^{n-k} \\
&= \frac{\lambda^k}{k!}\cdot \lim_{n\to\infty}\left(\left[\frac{n(n-1)(n-2)\dots(n-k+1)}{n^k}\right]\left[\left(1-\frac{\lambda}{n}\right)^n\right]\left[\left(1-\frac{\lambda}{n}\right)^{-k}\right]\right).
\end{aligned}$$

Note that for a fixed k, we have

$$\begin{aligned}
\lim_{n\to\infty} \frac{n(n-1)(n-2)\dots(n-k+1)}{n^k} &= 1 \\
\lim_{n\to\infty} \left(1-\frac{\lambda}{n}\right)^{-k} &= 1 \\
\lim_{n\to\infty} \left(1-\frac{\lambda}{n}\right)^{n} &= e^{-\lambda}.
\end{aligned}$$

Thus, we conclude

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

Discrete Random Variables (Cont.)

Cumulative Distribution Function

- The PMF is one way to describe the distribution of a discrete random variable. 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 random variable (discrete, continuous, and mixed).

The cumulative distribution function (CDF) of random variable X is defined as

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

Example 1:

I toss a coin twice. Let X be the number of observed heads. Find the CDF of X.

Note that here $X \sim Binomial(2, \frac{1}{2})$. The range of X is $R_X = \{0, 1, 2\}$ and its PMF is given by

$$P_X(0) = P(X = 0) = \frac{1}{4},$$

$$P_X(1) = P(X = 1) = \frac{1}{2},$$

$$P_X(2) = P(X = 2) = \frac{1}{4}.$$

To find the CDF, we argue as follows. First, note that if $x < 0$, then

$$F_X(x) = P(X \leq x) = 0, \text{ for } x < 0.$$

Next, if $x \geq 2$,

$$F_X(x) = P(X \leq x) = 1, \text{ for } x \geq 2.$$

Next, if $0 \leq x < 1$,

$$F_X(x) = P(X \leq x) = P(X = 0) = \frac{1}{4}, \text{ for } 0 \leq x < 1.$$

Finally, if $1 \leq x < 2$,

$$F_X(x) = P(X \leq x) = P(X = 0) + P(X = 1) = \frac{1}{4} + \frac{1}{2} = \frac{3}{4}, \text{ for } 1 \leq x < 2.$$

Thus, to summarize, we have

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

Example 2:

Let X be a discrete random variable with range $R_X = \{1, 2, 3, \ldots\}$. Suppose the PMF of X given by

$$P_X(k) = \frac{1}{2^k} \text{ for } k = 1, 2, 3, \ldots$$

a. Find and plot the CDF of X, $F_X(x)$.
b. Find $P(2 < X \leq 5)$.
c. Find $P(X > 4)$.

First, note that this is a valid PMF. In particular,

$$\sum_{k=1}^{\infty} P_X(k) = \sum_{k=1}^{\infty} \frac{1}{2^k} = 1 \text{ (geometric sum)}$$

a. To find the CDF, note that

$$\text{For } x < 1, \quad F_X(x) = 0.$$
$$\text{For } 1 \leq x < 2, F_X(x) = P_X(1) = \tfrac{1}{2}.$$
$$\text{For } 2 \leq x < 3, F_X(x) = P_X(1) + P_X(2) = \tfrac{1}{2} + \tfrac{1}{4} = \tfrac{3}{4}.$$

In general we have

$$\text{For } 0 < k \leq x < k + 1,$$
$$F_X(x) = P_X(1) + P_X(2) + \ldots + P_X(k)$$
$$= \frac{1}{2} + \frac{1}{4} + \ldots + \frac{1}{2^k} = \frac{2^k - 1}{2^k}.$$

b. To find $P(2 < X \leq 5)$, we can write

$$P(2 < X \leq 5) = F_X(5) - F_X(2) = \frac{31}{32} - \frac{3}{4} = \frac{7}{32}.$$

Or equivalently, we can write

$$P(2 < X \leq 5) = P_X(3) + P_X(4) + P_X(5) = \frac{1}{8} + \frac{1}{16} + \frac{1}{32} = \frac{7}{32},$$

which gives the same answer.

c. To find $P(X > 4)$, we can write

$$P(X > 4) = 1 - P(X \leq 4) = 1 - F_X(4) = 1 - \frac{15}{16} = \frac{1}{16}.$$

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.

Expected value (= mean=average):

Let X be a discrete random variable with range $R_X = \{x_1, x_2, x_3, \ldots\}$ (finite or countably infinite). The *expected* value of X, denoted by EX is defined as

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

To understand the concept behind EX, consider a discrete random variable with range $R_X = \{x_1, x_2, x_3, \dots\}$. This random variable is a result of random experiment. Suppose that we repeat this experiment a very large number of times N, and that the trials are independent. Let N_1 be the number of times we observe x_1, N_2 be the number of times we observe x_2, ..., N_k be the number of times we observe x_k, and so on. Since $P(X = x_k) = P_X(x_k)$, we expect that

$$P_X(x_1) \approx \frac{N_1}{N},$$

$$P_X(x_2) \approx \frac{N_2}{N},$$

$$\cdot \quad \cdot \quad \cdot$$

$$P_X(x_k) \approx \frac{N_k}{N},$$

$$\cdot \quad \cdot \quad \cdot$$

In other words, we have $N_k \approx N P_X(x_k)$. Now, if we take the average of the observed values of X, we obtain

In other words, we have $N_k \approx NP_X(x_k)$. Now, if we take the average of the observed values of X, we obtain

$$\begin{aligned}\text{Average} &= \frac{N_1x_1+N_2x_2+N_3x_3+\ldots}{N}\\ &\approx \frac{x_1NP_X(x_1)+x_2NP_X(x_2)+x_3NP_X(x_3)+\ldots}{N}\\ &= x_1P_X(x_1)+x_2P_X(x_2)+x_3P_X(x_3)+\ldots\\ &= EX.\end{aligned}$$

Thus, the intuition behind EX is that if you repeat the random experiment independently N times and take the average of the observed data, the average gets closer and closer to EX as N gets larger and larger. We sometimes denote EX by μ_X.

Different notations for expected value of X: $EX = E[X] = E(X) = \mu_X$.

Let's compute the expected values of some well-known distributions.

Example 3: Let X~Bernoulli(p). Find EX.

For the Bernoulli distribution, the range of X is $R_X = \{0, 1\}$, and $P_X(1) = p$ and $P_X(0) = 1 - p$. Thus,

$$\begin{aligned} EX &= 0 \cdot P_X(0) + 1 \cdot P_X(1) \\ &= 0 \cdot (1-p) + 1 \cdot p \\ &= p. \end{aligned}$$

Linearity

An important property of expectation, which is *linearity*. Note that if X is a random variable, any function of X is also a random variable, so we can talk about its expected value. For example, if Y=aX+b, we can talk about EY=E[aX+b]. Or if you define $Y=X_1+X_2+\cdots+X_n$, where X_is are random variables, we can talk about $EY=E[X_1+X_2+\cdots+X_n]$. The following theorem states that expectation is linear, which makes it easier to calculate the expected value of linear functions of random variables.

Expectation is linear:

Theorem

We have

- $E[aX+b]=aEX+b$, for all $a,b\in\mathbb{R}$;
- $E[X_1+X_2+\cdots+X_n]=EX_1+EX_2+\cdots+EX_n$, for any set of random variables $X_1,X_2,\cdots,X_n$.

Functions of Random Variables

If X is a random variable 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 4:

Let X be a discrete random variable with $P_X(k)=15$ for $k=-1,0,1,2,3$. Let $Y=2|X|$. Find the range and PMF of Y

First, note that the range of Y is

$$\begin{aligned} R_Y &= \{2|x| \text{ where } x \in R_X\} \\ &= \{0, 2, 4, 6\}. \end{aligned}$$

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

$$\begin{aligned} P_Y(0) &= P(Y = 0) = P(2|X| = 0) \\ &= P(X = 0) = \tfrac{1}{5}; \\ P_Y(2) &= P(Y = 2) = P(2|X| = 2) \\ &= P((X = -1) \text{ or } (X = 1)) \\ &= P_X(-1) + P_X(1) = \tfrac{1}{5} + \tfrac{1}{5} = \tfrac{2}{5}; \\ P_Y(4) &= P(Y = 4) = P(2|X| = 4) \\ &= P(X = 2) + P(X = -2) = \tfrac{1}{5}; \\ P_Y(6) &= P(Y = 6) = P(2|X| = 6) \\ &= P(X = 3) + P(X = -3) = \tfrac{1}{5}. \end{aligned}$$

So, to summarize,

$$P_Y(k) = \begin{cases} \frac{1}{5} & \text{for } k = 0, 4, 6 \\ \frac{2}{5} & \text{for } k = 2 \\ 0 & \text{otherwise} \end{cases}$$

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 $EY=E[g(X)]=\sum_{y\in R_Y} yP_Y(y)$. But there is another way which is usually easier. It is called the law of the unconscious statistician (LOTUS).

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

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

Example

Let X be a discrete random variable with range $R_X = \{0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}, \pi\}$, such that $P_X(0) = P_X(\frac{\pi}{4}) = P_X(\frac{\pi}{2}) = P_X(\frac{3\pi}{4}) = P_X(\pi) = \frac{1}{5}$. Find $E[\sin(X)]$.

Using LOTUS, we have

$$\begin{aligned} E[g(X)] &= \sum_{x_k \in R_X} g(x_k) P_X(x_k) \\ &= \sin(0) \cdot \tfrac{1}{5} + \sin(\tfrac{\pi}{4}) \cdot \tfrac{1}{5} + \sin(\tfrac{\pi}{2}) \cdot \tfrac{1}{5} + \sin(\tfrac{3\pi}{4}) \cdot \tfrac{1}{5} + \sin(\pi) \cdot \tfrac{1}{5} \\ &= 0 \cdot \tfrac{1}{5} + \tfrac{\sqrt{2}}{2} \cdot \tfrac{1}{5} + 1 \cdot \tfrac{1}{5} + \tfrac{\sqrt{2}}{2} \cdot \tfrac{1}{5} + 0 \cdot \tfrac{1}{5} \\ &= \tfrac{\sqrt{2}+1}{5}. \end{aligned}$$

Variance

Consider two random variables X and Y with the following PMFs.

$$P_X(x) = \begin{cases} 0.5 & \text{for } x = -100 \\ 0.5 & \text{for } x = 100 \\ 0 & \text{otherwise} \end{cases}$$

$$P_Y(y) = \begin{cases} 1 & \text{for } y = 0 \\ 0 & \text{otherwise} \end{cases}$$

Note that $EX = EY = 0$. Although both random variables have the same mean value, their distribution is completely different. Y is always equal to its mean of 0, while X is either 100 or -100, quite far from its mean value. The **variance** is a measure of how spread out the distribution of a random variable is. Here, the variance of Y is quite small since its distribution is concentrated at a single value, while the variance of X will be larger since its distribution is more spread out.

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

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

By definition, the variance of X is the average value of $(X-\mu_X)^2$. Since $(X-\mu_X)^2 \geq 0$, the variance is always larger than or equal to zero. A large value of the variance means that $(X-\mu_X)^2$ is often large, so X often takes values far from its mean. This means that the distribution is very spread out. On the other hand, a low variance means that the distribution is concentrated around its average.

Note that if we did not square the difference between X and its mean, the result would be 0. That is

$$E[X-\mu_X] = EX - E[\mu_X] = \mu_X - \mu_X = 0.$$

X is sometimes below its average and sometimes above its average. Thus, $X-\mu_X$ is sometimes negative and sometimes positive, but on average it is zero.

To compute $Var(X) = E\big[(X-\mu_X)^2\big]$, 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

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

The **standard deviation** of a random variable X is defined as

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

Computational formula for the variance:

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

Example

I roll a fair die and let X be the resulting number. Find EX, $\text{Var}(X)$, and σ_X.

We have $R_X = \{1,2,3,4,5,6\}$ and $P_X(k) = \frac{1}{6}$ for $k = 1,2,\dots,6$. Thus, we have

$$EX = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = \frac{7}{2};$$

$$EX^2 = 1 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 9 \cdot \frac{1}{6} + 16 \cdot \frac{1}{6} + 25 \cdot \frac{1}{6} + 36 \cdot \frac{1}{6} = \frac{91}{6}.$$

Thus

$$\text{Var}(X) = E[X^2] - (EX)^2 = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{91}{6} - \frac{49}{4} \approx 2.92,$$

$$\sigma_X = \sqrt{\text{Var}(X)} \approx \sqrt{2.92} \approx 1.71$$

Theorem

For a random variable X and real numbers a and b,

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

Proof

If $Y = aX + b$, $EY = aEX + b$. Thus,

$$\begin{aligned}\mathrm{Var}(Y) &= E[(Y-EY)^2] \\ &= E[(aX+b-aEX-b)^2] \\ &= E[a^2(X-\mu_X)^2] \\ &= a^2E[(X-\mu_X)^2] \\ &= a^2\mathrm{Var}(X)\end{aligned}$$

Theorem .

If $X_1, X_2, \cdots, X_n$ are independent random variables and $X = X_1 + X_2 + \cdots + X_n$, then

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

Example

If $X \sim Binomial(n, p)$ find Var(X).

We know that we can write a $Binomial(n, p)$ random variable as the sum of n **independent** $Bernoulli(p)$ random variables, i.e., $X = X_1 + X_2 + \cdots + X_n$. Thus, we conclude

$$\mathbf{Var}(X) = \mathbf{Var}(X_1) + \mathbf{Var}(X_2) + \cdots + \mathbf{Var}(X_n).$$

If $X_i \sim Bernoulli(p)$, then its variance is

$$\mathbf{Var}(X_i) = E[X_i^2] - (EX_i)^2 = 1^2 \cdot p + 0^2 \cdot (1-p) - p^2 = p(1-p).$$

Thus,

$$\begin{aligned} \mathbf{Var}(X) &= p(1-p) + p(1-p) + \cdots + p(1-p) \\ &= np(1-p). \end{aligned}$$

JOINT DISTRIBUTION

Two Random Variables

- 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. 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.

Joint Probability Mass Function (PMF)

- A discrete random variable X, we define the PMF as $P_X(x)=P(X=x)$. Now, if we have two random variables X and Y, and we would like to study them jointly, we define the **joint probability mass function** as follows:

The **joint probability mass function** of two discrete random variables X and Y is defined as

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

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

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

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 we can always write

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

In fact, sometimes we define $R_{XY} = R_X \times R_Y$ to simplify the analysis. In this case, for some pairs (x_i, y_j) in $R_X \times R_Y$, $P_{XY}(x_i, y_j)$ might be zero. For two discrete random variables X and Y, 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 $P\big((X,Y) \in A\big)$ for any set $A \subset \mathbb{R}^2$. Specifically, we have

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

Note that the event $X = x$ can be written as $\{(x_i, y_j) : x_i = x, y_j \in R_Y\}$. Also, the event $Y = y$ can be written as $\{(x_i, y_j) : x_i \in R_X, y_j = y\}$. Thus, we can write

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

Marginal PMF

The joint PMF contains all the information regarding the distributions of X and Y. This means that, for example, we can obtain PMF of X from its joint PMF with Y. Indeed, we can write

$$\begin{aligned} P_X(x) &= P(X=x) \\ &= \sum_{y_j \in R_Y} P(X=x, Y=y_j) \qquad \text{law of total probablity} \\ &= \sum_{y_j \in R_Y} P_{XY}(x, y_j). \end{aligned}$$

Here, we call $P_X(x)$ the **marginal PMF** of X. Similarly, we can find the marginal PMF of Y as

$$P_Y(Y) = \sum_{x_i \in R_X} P_{XY}(x_i, y).$$

Marginal PMFs of X and Y:

$$P_X(x) = \sum_{y_j \in R_Y} P_{XY}(x, y_j), \qquad \text{for any } x \in R_X$$

$$P_Y(y) = \sum_{x_i \in R_X} P_{XY}(x_i, y), \qquad \text{for any } y \in R_Y$$

Example

Consider two random variables X and Y with joint PMF given in Table 1.

Table 1 Joint PMF of X and Y in Example 1

	$Y=0$	$Y=1$	$Y=2$
$X=0$	$\frac{1}{6}$	$\frac{1}{4}$	$\frac{1}{8}$
$X=1$	$\frac{1}{8}$	$\frac{1}{6}$	$\frac{1}{6}$

Figure shows $P_{XY}(x,y)$.

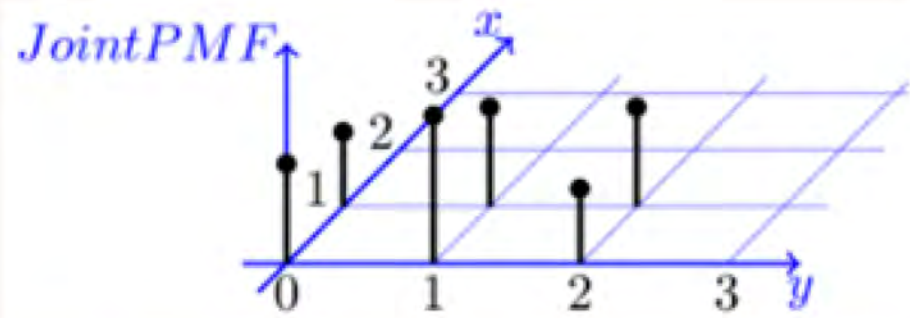


Figure : Joint PMF of X and Y

a. Find $P(X=0, Y \leq 1)$.
b. Find the marginal PMFs of X and Y.
c. Find $P(Y=1|X=0)$.
d. Are X and Y independent?

a. To find $P(X = 0, Y \leq 1)$, we can write

$$P(X = 0, Y \leq 1) = P_{XY}(0,0) + P_{XY}(0,1) = \frac{1}{6} + \frac{1}{4} = \frac{5}{12}.$$

b. Note that from the table,

$$R_X = \{0, 1\} \qquad \text{and} \qquad R_Y = \{0, 1, 2\}.$$

Now we can use Equation 5.1 to find the marginal PMFs. For example, to find $P_X(0)$, we can write

$$\begin{aligned} P_X(0) &= P_{XY}(0,0) + P_{XY}(0,1) + P_{XY}(0,2) \\ &= \frac{1}{6} + \frac{1}{4} + \frac{1}{8} \\ &= \frac{13}{24}. \end{aligned}$$

We obtain

$$P_X(x) = \begin{cases} \frac{13}{24} & x = 0 \\ \frac{11}{24} & x = 1 \\ 0 & \text{otherwise} \end{cases}$$

$$P_Y(y) = \begin{cases} \frac{7}{24} & y = 0 \\ \frac{5}{12} & y = 1 \\ \frac{7}{24} & y = 2 \\ 0 & \text{otherwise} \end{cases}$$

c. Find $P(Y = 1|X = 0)$: Using the formula for conditional probability, we have

$$\begin{aligned} P(Y = 1|X = 0) &= \frac{P(X = 0, Y = 1)}{P(X = 0)} \\ &= \frac{P_{XY}(0, 1)}{P_X(0)} \\ &= \frac{\frac{1}{4}}{\frac{13}{24}} = \frac{6}{13}. \end{aligned}$$

d. Are X and Y independent? X and Y are not independent, because as we just found out

$$P(Y = 1|X = 0) = \frac{6}{13} \neq P(Y = 1) = \frac{5}{12}.$$

Joint Cumulative Distributive Function (CDF)

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

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

As usual, comma means "and," so we can write

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

Below figure shows the region associated with $F_{XY(x,y)}$ in the two-dimensional plane. Note that the above definition of joint CDF is a general ddefinition and is applicable to discrete, continuous, and mixed random variables. Since the joint CDF refers to the probability of an event, we must have $0 \leq F_{XY}(x,y) \leq 1$

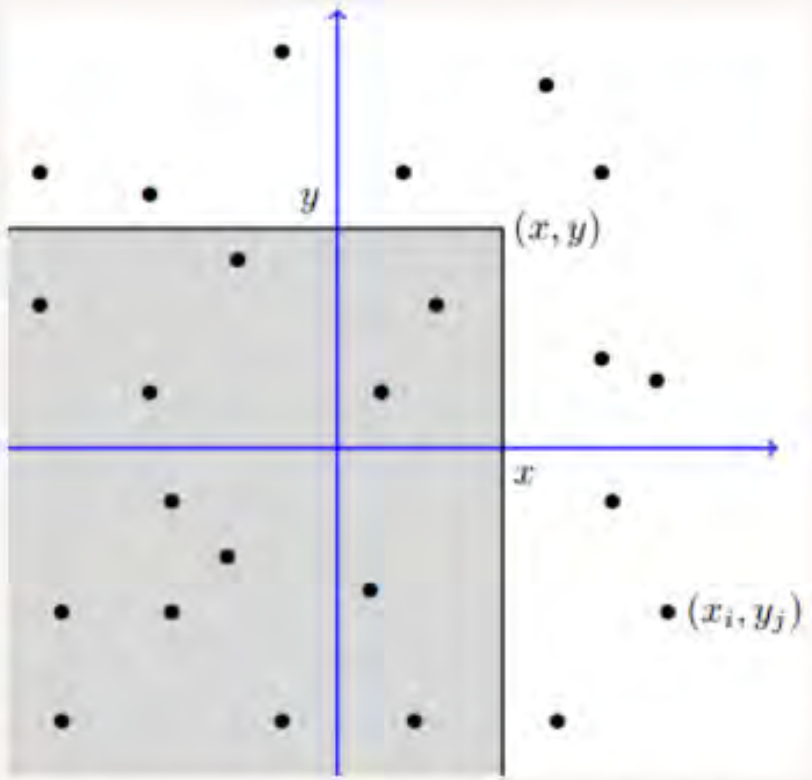


Figure $F_{XY}(x,y)$ is the probability that (X,Y) belongs to the shaded region. The dots are the pairs (x_i, y_j) in R_{XY}.

If we know the joint 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

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

Marginal CDF

Marginal CDFs of X and Y:

$$F_X(x) = F_{XY}(x, \infty) = \lim_{y \to \infty} F_{XY}(x, y), \qquad \text{for any } x,$$
$$F_Y(y) = F_{XY}(\infty, y) = \lim_{x \to \infty} F_{XY}(x, y), \qquad \text{for any } y \qquad (5.2)$$

Also, note that we must have

$$F_{XY}(\infty, \infty) = 1,$$
$$F_{XY}(-\infty, y) = 0, \qquad \text{for any } y,$$
$$F_{XY}(x, -\infty) = 0, \qquad \text{for any } x.$$

Conditioning and Independence

We have discussed conditional probability before, and you have already seen some problems regarding random variables and conditional probability. Here, we will discuss conditioning for random variables more in detail and introduce the conditional PMF, conditional CDF, and conditional expectation. We would like to emphasize that there is only one main formula regarding conditional probability which is

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

Any other formula regarding conditional probability can be derived from the above formula. Specifically, if you have two random variables X and Y, you can write

$$P(X \in C|Y \in D) = \frac{P(X \in C, Y \in D)}{P(Y \in D)}, \text{ where } C, D \subset \mathbb{R}.$$

For a discrete random variable X and event A, the **conditional PMF** of X given A is defined as

$$\begin{aligned} P_{X|A}(x_i) &= P(X = x_i|A) \\ &= \frac{P(X = x_i \text{ and } A)}{P(A)}, \quad \text{for any } x_i \in R_X. \end{aligned}$$

Similarly, we define the **conditional CDF** of X given A as

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

Conditional PMF of X Given Y:

In some problems, 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. The conditional PMF of X given Y is defined as

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

Similarly, we can define the conditional probability of Y given X:

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

For discrete random variables X and Y, the **conditional PMFs** of X given Y and vice versa are defined as

$$P_{X|Y}(x_i|y_j) = \frac{P_{XY}(x_i,y_j)}{P_Y(y_j)},$$

$$P_{Y|X}(y_j|x_i) = \frac{P_{XY}(x_i,y_j)}{P_X(x_i)}$$

for any $x_i \in R_X$ and $y_j \in R_Y$.

Independent Random Variables:

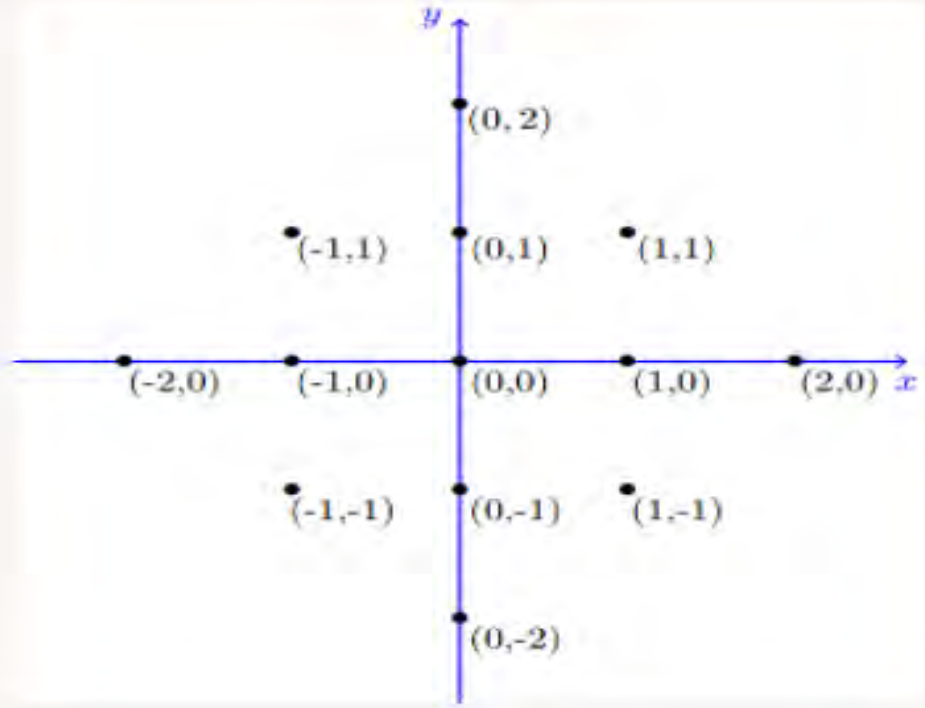
We have defined independent random variables previously. Now that we have seen joint PMFs and CDFs, we can restate the independence definition.

Two discrete random variables X and Y are independent if

$$P_{XY}(x, y) = P_X(x)P_Y(y), \quad \text{for all } x, y.$$

Equivalently, X and Y are independent if

$$F_{XY}(x, y) = F_X(x)F_Y(y), \quad \text{for all } x, y.$$



So, if X and Y are independent, we have

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

As we expect, for independent random variables, the conditional PMF is equal to the marginal PMF. In other words, knowing the value of Y does not provide any information about X.

Conditional Expectation:

Given that we know event A has occurred, we can compute the conditional expectation of a random variable X, $E[X|A]$. Conditional expectation is similar to ordinary expectation. The only difference is that we replace the PMF by the conditional PMF. Specifically, we have

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

Similarly, given that we have observed the value of random variable Y, we can compute the conditional expectation of X. Specifically, the conditional expectation of X given that $Y = y$ is

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

Conditional Expectation of X:

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

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

Law of Total Probability:

Law of Total Probability:

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

Law of Total Expectation:

1. If $B_1, B_2, B_3, \ldots$ is a partition of the sample space S,

$$EX = \sum_i E[X|B_i] P(B_i)$$

2. For a random variable X and a discrete random variable Y,

$$EX = \sum_{y_j \in R_Y} E[X|Y = y_j] P_Y(y_j)$$

Functions of Two Random Variables

Analysis of a function of two random variables is pretty much the same as for a function of a single random variable. Suppose that you have two discrete random variables X and Y, and suppose that $Z = g(X, Y)$, where $g : \mathbb{R}^2 \mapsto \mathbb{R}$. Then, if we are interested in the PMF of Z, we can write

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

Note that if we are only interested in $E[g(X, Y)]$, we can directly use LOTUS, without finding $P_Z(z)$:

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

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

Example

Linearity of Expectation: For two discrete random variables X and Y, show that $E[X+Y] = EX + EY$.

Let $g(X,Y) = X + Y$. Using LOTUS, we have

$$
\begin{aligned}
E[X+Y] &= \sum_{(x_i,y_j)\in R_{XY}} (x_i + y_j) P_{XY}(x_i, y_j) \\
&= \sum_{(x_i,y_j)\in R_{XY}} x_i P_{XY}(x_i, y_j) + \sum_{(x_i,y_j)\in R_{XY}} y_j P_{XY}(x_i, y_j) \\
&= \sum_{x_i \in R_X} \sum_{y_j \in R_Y} x_i P_{XY}(x_i, y_j) + \sum_{x_i \in R_X} \sum_{y_j \in R_Y} y_j P_{XY}(x_i, y_j) \\
&= \sum_{x_i \in R_X} x_i \sum_{y_j \in R_Y} P_{XY}(x_i, y_j) + \sum_{y_j \in R_Y} y_j \sum_{x_i \in R_X} P_{XY}(x_i, y_j) \\
&= \sum_{x_i \in R_X} x_i P_X(x_i) + \sum_{y_j \in R_Y} y_j P_Y(y_j) \qquad \text{(marginal PMF (Equation 5.1))} \\
&= EX + EY.
\end{aligned}
$$

Conditional Expectation (Revisited) and Conditional Variance

- We briefly discussed conditional expectation. Here, we will discuss the properties of conditional expectation in more detail as they are quite useful in practice. We will also discuss conditional variance. An important concept here is that we interpret the conditional expectation as a random variable.

Conditional Expectation as a Function of a Random Variable:

Remember that the conditional expectation of X given that $Y = y$ is given by

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

Note that $E[X|Y = y]$ depends on the value of y. In other words, by changing y, $E[X|Y = y]$ can also change. Thus, we can say $E[X|Y = y]$ is a function of y, so let's write

$$g(y) = E[X|Y = y].$$

Thus, we can think of $g(y) = E[X|Y = y]$ as a function of the value of random variable Y. We then write

$$g(Y) = E[X|Y].$$

We use this notation to indicate that $E[X|Y]$ is a random variable whose value equals $g(y) = E[X|Y = y]$ when $Y = y$. Thus, if Y is a random variable with range $R_Y = \{y_1, y_2, \cdots\}$, then $E[X|Y]$ is also a random variable with

$$E[X|Y] = \begin{cases} E[X|Y = y_1] & \text{with probability } P(Y = y_1) \\ E[X|Y = y_2] & \text{with probability } P(Y = y_2) \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \end{cases}$$

Example . Consider two random variables X and Y with joint PMF given in Table 2 Let $Z = E[X|Y]$.

a. Find the Marginal PMFs of X and Y.
b. Find the conditional PMF of X given $Y = 0$ and $Y = 1$, i.e., find $P_{X|Y}(x|0)$ and $P_{X|Y}(x|1)$.
c. Find the PMF of Z.
d. Find EZ, and check that $EZ = EX$.
e. Find Var(Z).

Table 2: Joint PMF of X and Y in example .

	$Y = 0$	$Y = 1$
$X = 0$	$\frac{1}{5}$	$\frac{2}{5}$
$X = 1$	$\frac{2}{5}$	0

a. Using the table we find out

$$P_X(0) = \frac{1}{5} + \frac{2}{5} = \frac{3}{5},$$
$$P_X(1) = \frac{2}{5} + 0 = \frac{2}{5},$$
$$P_Y(0) = \frac{1}{5} + \frac{2}{5} = \frac{3}{5},$$
$$P_Y(1) = \frac{2}{5} + 0 = \frac{2}{5}.$$

Thus, the marginal distributions of X and Y are both $Bernoulli(\frac{2}{5})$. However, note that X and Y are not independent.

b. We have

$$P_{X|Y}(0|0) = \frac{P_{XY}(0,0)}{P_Y(0)}$$
$$= \frac{\frac{1}{5}}{\frac{3}{5}} = \frac{1}{3}.$$

Thus,

$$P_{X|Y}(1|0) = 1 - \frac{1}{3} = \frac{2}{3}.$$

We conclude

$$X|Y=0 \sim Bernoulli\left(\frac{2}{3}\right).$$

Similarly, we find

$$P_{X|Y}(0|1) = 1,$$
$$P_{X|Y}(1|1) = 0.$$

Thus, given $Y = 1$, we have always $X = 0$.

c. We note that the random variable Y can take two values: 0 and 1. Thus, the random variable $Z = E[X|Y]$ can take two values as it is a function of Y. Specifically,

$$Z = E[X|Y] = \begin{cases} E[X|Y=0] & \text{if } Y = 0 \\ \\ E[X|Y=1] & \text{if } Y = 1 \end{cases}$$

Now, using the previous part, we have

$$E[X|Y=0] = \frac{2}{3}, \qquad E[X|Y=1] = 0,$$

and since $P(y=0) = \frac{3}{5}$, and $P(y=1) = \frac{2}{5}$, we conclude that

$$Z = E[X|Y] = \begin{cases} \frac{2}{3} & \text{with probability } \frac{3}{5} \\ \\ 0 & \text{with probability } \frac{2}{5} \end{cases}$$

So we can write

$$P_Z(z) = \begin{cases} \frac{3}{5} & \text{if } z = \frac{2}{3} \\ \\ \frac{2}{5} & \text{if } z = 0 \\ \\ 0 & \text{otherwise} \end{cases}$$

d. Now that we have found the PMF of Z, we can find its mean and variance. Specifically,

$$E[Z] = \frac{2}{3} \cdot \frac{3}{5} + 0 \cdot \frac{2}{5} = \frac{2}{5}.$$

We also note that $EX = \frac{2}{5}$. Thus, here we have

$$E[X] = E[Z] = E[E[X|Y]].$$

In fact, as we will prove shortly, the above equality always holds. It is called the law of iterated expectations.

e. To find Var(Z), we write

$$\begin{aligned} \mathbf{Var}(Z) &= E[Z^2] - (EZ)^2 \\ &= E[Z^2] - \frac{4}{25}, \end{aligned}$$

where

$$E[Z^2] = \frac{4}{9} \cdot \frac{3}{5} + 0 \cdot \frac{2}{5} = \frac{4}{15}.$$

Thus,

$$\begin{aligned} \mathbf{Var}(Z) &= \frac{4}{15} - \frac{4}{25} \\ &= \frac{8}{75}. \end{aligned}$$

Expectation for Independent Random Variables:

Note that if two random variables X and Y are independent, then the conditional PMF of X given Y will be the same as the marginal PMF of X, i.e., for any $x \in R_X$, we have

$$P_{X|Y}(x|y) = P_X(x).$$

Thus, for independent random variables, we have

$$\begin{aligned} E[X|Y=y] &= \sum_{x \in R_X} x P_{X|Y}(x|y) \\ &= \sum_{x \in R_X} x P_X(x) \\ &= E[X]. \end{aligned}$$

Again, thinking of this as a random variable depending on Y, we obtain

$$E[X|Y] = E[X], \text{ when } X \text{ and } Y \text{ are independent.}$$

More generally, if X and Y are independent then any function of X, say $g(X)$, and Y are independent, thus

$$E[g(X)|Y] = E[g(X)].$$

Remember that for independent random variables, $P_{XY}(x,y) = P_X(x)P_Y(y)$. From this, we can show that $E[XY] = EXEY$.

If X and Y are independent random variables, then

1. $E[X|Y] = EX$;
2. $E[g(X)|Y] = E[g(X)]$;
3. $E[XY] = EXEY$;
4. $E[g(X)h(Y)] = E[g(X)]E[h(Y)]$.

Conditional Variance:

Similar to the conditional expectation, we can define the conditional variance of X, $\text{Var}(X|Y=y)$, which is the variance of X in the conditional space where we know $Y=y$. If we let $\mu_{X|Y}(y) = E[X|Y=y]$, then

$$\begin{aligned}\text{Var}(X|Y=y) &= E\left[(X-\mu_{X|Y}(y))^2|Y=y\right] \\ &= \sum_{x_i \in R_X} \left(x_i-\mu_{X|Y}(y)\right)^2 P_{X|Y}(x_i) \\ &= E\left[X^2|Y=y\right]-\mu_{X|Y}(y)^2.\end{aligned}$$

Note that $\text{Var}(X|Y=y)$ is a function of y. Similar to our discussion on $E[X|Y=y]$ and $E[X|Y]$, we define $\text{Var}(X|Y)$ as a function of the random variable Y. That is, $\text{Var}(X|Y)$ is a random variable whose value equals $\text{Var}(X|Y=y)$ whenever $Y=y$

Law of Total Variance:

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

Continuous and Mixed Random Variables

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 $x \in R$, $P(X=x)=0$. 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. Of course, there is a little bit more to the story and that's why we need a chapter to discuss it. In this chapter, we will also introduce mixed random variables that are mixtures of discrete and continuous random variables.

Continuous Random Variables and their Distributions

A random variable X with CDF $F_X(x)$ is said to be continuous if $F_X(x)$ is a continuous function for all $x \in \mathbb{R}$.

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 $P(X=x)=0$ for all $x \in R$. 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. 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

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

So, we conclude that

$$f_X(x) = \lim_{\Delta \to 0} \frac{F_X(x + \Delta) - F_X(x)}{\Delta}$$

$$= \frac{dF_X(x)}{dx} = F'_X(x), \qquad \text{if } F_X(x) \text{ is differentiable at } x.$$

Thus, we have the following definition for the PDF of continuous random variables

Consider a continuous random variable X with an absolutely continuous CDF $F_X(x)$. The function $f_X(x)$ defined by

$$f_X(x) = \frac{dF_X(x)}{dx} = F'_X(x), \qquad \text{if } F_X(x) \text{ is differentiable at } x$$

is called the probability density function (PDF) of X.

Consider a continuous random variable X with PDF $f_X(x)$. We have

1. $f_X(x) \geq 0$ for all $x \in \mathbb{R}$.
2. $\int_{-\infty}^{\infty} f_X(u)du = 1$.
3. $P(a < X \leq b) = F_X(b) - F_X(a) = \int_a^b f_X(u)du$.
4. More generally, for a set A, $P(X \in A) = \int_A f_X(u)du$.

Let X be a continuous random variable with the following PDF

$$f_X(x) = \begin{cases} ce^{-x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

where c is a positive constant.

a. Find c.
b. Find the CDF of X, $F_X(x)$.
c. Find $P(1 < X < 3)$.

a. To find c, we can use Property 2 above, in particular

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty} f_X(u)du \\ &= \int_0^{\infty} ce^{-u}du \\ &= c\Big[-e^{-x}\Big]_0^{\infty} \\ &= c. \end{aligned}$$

Thus, we must have $c = 1$.

b. To find the CDF of X, we use $F_X(x) = \int_{-\infty}^{x} f_X(u)du$, so for $x < 0$, we obtain $F_X(x) = 0$. For $x \geq 0$, we have

$$F_X(x) = \int_0^x e^{-u}du = 1 - e^{-x}.$$

Thus,

$$F_X(x) = \begin{cases} 1 - e^{-x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

c. We can find $P(1 < X < 3)$ using either the CDF or the PDF. If we use the CDF, we have

$$P(1 < X < 3) = F_X(3) - F_X(1) = \left[1 - e^{-3}\right] - \left[1 - e^{-1}\right] = e^{-1} - e^{-3}.$$

Equivalently, we can use the PDF. We have

$$\begin{aligned} P(1 < X < 3) &= \int_1^3 f_X(t)dt \\ &= \int_1^3 e^{-t}dt \\ &= e^{-1} - e^{-3}. \end{aligned}$$

Expected Value and Variance

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

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

Now, by replacing the sum by an integral and PMF by PDF, we can write the definition of expected value of a continuous random variable as

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

Example

Let $X \sim Uniform(a,b)$. Find EX.

As we saw, the PDF of X is given by

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

so to find its expected value, we can write

$$\begin{aligned} EX &= \int_{-\infty}^{\infty} x f_X(x) dx \\ &= \int_a^b x(\frac{1}{b-a}) dx \\ &= \frac{1}{b-a}\left[\frac{1}{2}x^2\right]_a^b dx \\ &= \frac{a+b}{2}. \end{aligned}$$

This result is intuitively reasonable: since X is uniformly distributed over the interval $[a,b]$, we expect its mean to be the middle point, i.e., $EX = \frac{a+b}{2}$.

Example

Let X be a continuous random variable with PDF

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

Find the expected value of X.

We have

$$\begin{aligned} EX &= \int_{-\infty}^{\infty} x f_X(x) dx \\ &= \int_0^1 x(2x) dx \\ &= \int_0^1 2x^2 dx \\ &= \frac{2}{3}. \end{aligned}$$

Expected Value of a Function of a Continuous Random Variable

Remember the law of the unconscious statistician (LOTUS) for discrete random variables:

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

Now, by changing the sum to integral and changing the PMF to PDF we will obtain the similar formula for continuous random variables.

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

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

As we have seen before, expectation is a linear operation, thus we always have

- $E[aX + b] = aEX + b$, for all $a, b \in \mathbb{R}$, and
- $E[X_1 + X_2 + \ldots + X_n] = EX_1 + EX_2 + \ldots + EX_n$, for any set of random variables $X_1, X_2, \ldots, X_n$.

Example

Let X be a continuous random variable with PDF

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

Find $E(X^n)$, where $n \in \mathbb{N}$.

Using LOTUS we have

$$\begin{aligned} E[X^n] &= \int_{-\infty}^{\infty} x^n f_X(x) dx \\ &= \int_0^1 x^n (x + \tfrac{1}{2}) dx \\ &= \left[\frac{1}{n+2} x^{n+2} + \frac{1}{2(n+1)} x^{n+1} \right]_0^1 \\ &= \frac{3n+4}{2(n+1)(n+2)}. \end{aligned}$$

Variance

Remember that the variance of any random variable is defined as

$$\mathrm{Var}(X) = E\left[(X-\mu_X)^2\right] = EX^2 - (EX)^2.$$

So for a continuous random variable, we can write

$$\mathrm{Var}(X) = E\left[(X-\mu_X)^2\right] = \int_{-\infty}^{\infty} (x-\mu_X)^2 f_X(x)dx$$

$$= EX^2 - (EX)^2 = \int_{-\infty}^{\infty} x^2 f_X(x)dx - \mu_X^2$$

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

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

Example

Let X be a continuous random variable with PDF

$$f_X(x) = \begin{cases} \frac{3}{x^4} & x \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

Find the mean and variance of X.

$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} x f_X(x) dx \\ &= \int_1^{\infty} \frac{3}{x^3} dx \\ &= \left[-\frac{3}{2} x^{-2} \right]_1^{\infty} \\ &= \frac{3}{2}. \end{aligned}$$

Next, we find EX^2 using LOTUS,

$$\begin{aligned} E[X^2] &= \int_{-\infty}^{\infty} x^2 f_X(x) dx \\ &= \int_1^{\infty} \frac{3}{x^2} dx \\ &= \left[-3x^{-1} \right]_1^{\infty} \\ &= 3. \end{aligned}$$

Thus, we have

$$\mathrm{Var}(X) = EX^2 - (EX)^2 = 3 - \frac{9}{4} = \frac{3}{4}.$$

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. Let's look at an example.

Example

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

a. Find the CDF of Y.
b. Find the PDF of Y.
c. Find EY.

First, note that we already know the CDF and PDF of X. In particular,

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

Special Distributions

- Uniform Distributions
- Exponential Distribution
- Normal (Gaussian) Distribution
- Gamma Distribution

Uniform Distribution

A continuous random variable X is said to have a *Uniform* distribution over the interval $[a, b]$, shown as $X \sim Uniform(a, b)$, if its PDF is given by

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

For X~Uniform(a,b) mean wil be given as;

$$EX = \frac{a+b}{2}$$

To find the variance, we can find EX^2 using LOTUS:

$$\begin{aligned} EX^2 &= \int_{-\infty}^{\infty} x^2 f_X(x) dx \\ &= \int_a^b x^2 \left(\frac{1}{b-a}\right) dx \\ &= \frac{a^2+ab+b^2}{3}. \end{aligned}$$

Therefore,

$$\begin{aligned} \mathrm{Var}(X) &= EX^2 - (EX)^2 \\ &= \frac{(b-a)^2}{12}. \end{aligned}$$

Exponential distribution

- The exponential distribution is one of the widely used continuous distributions. It is often used to model the time elapsed between events. We will now mathematically define the exponential distribution, and derive its mean and expected value. Then we will develop the intuition for the distribution and discuss several interesting properties that it has.

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

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

It is convenient to use the unit step function defined as

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

so we can write the PDF of an $Exponential(\lambda)$ random variable as

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

Let us find its CDF, mean and variance. For $x > 0$, we have

$$F_X(x) = \int_0^x \lambda e^{-\lambda t} dt = 1 - e^{-\lambda x}.$$

So we can express the CDF as

$$F_X(x) = \left(1 - e^{-\lambda x}\right) u(x).$$

Let $X \sim Exponential(\lambda)$. We can find its expected value as follows, using integration by parts:

$$\begin{aligned} EX &= \int_0^\infty x \lambda e^{-\lambda x} dx \\ &= \frac{1}{\lambda} \int_0^\infty y e^{-y} dy \qquad \text{choosing } y = \lambda x \\ &= \frac{1}{\lambda} \Big[-e^{-y} - y e^{-y} \Big]_0^\infty \\ &= \frac{1}{\lambda}. \end{aligned}$$

Now let's find Var(X). We have

$$\begin{aligned} EX^2 &= \int_0^\infty x^2 \lambda e^{-\lambda x} dx \\ &= \frac{1}{\lambda^2} \int_0^\infty y^2 e^{-y} dy \\ &= \frac{1}{\lambda^2} \Big[-2e^{-y} - 2ye^{-y} - y^2 e^{-y} \Big]_0^\infty \\ &= \frac{2}{\lambda^2}. \end{aligned}$$

Thus, we obtain

$$\mathrm{Var}(X) = EX^2 - (EX)^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

If $X \sim Exponential(\lambda)$, then $EX = \frac{1}{\lambda}$ and Var$(X) = \frac{1}{\lambda^2}$.

Normal (Gaussian) Distribution

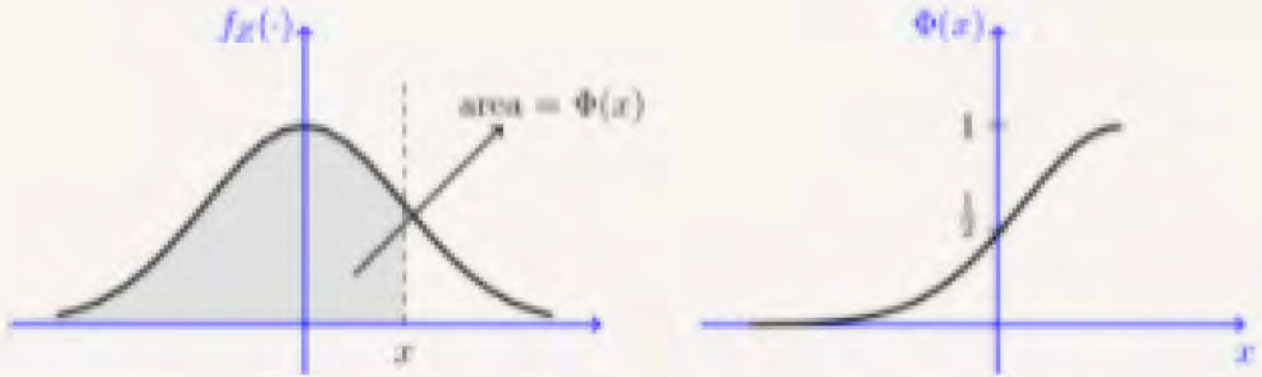
A continuous random variable Z is said to be a *standard normal (standard Gaussian)* random variable, shown as $Z \sim N(0,1)$, if its PDF is given by

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

If $Z \sim N(0,1)$, then $EZ = 0$ and $\text{Var}(Z) = 1$.

The CDF of the standard normal distribution is denoted by the Φ function:

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



The Φ function (CDF of standard normal).

Here are some properties of the Φ function that can be shown from its definition.

1. $\lim_{x \to \infty} \Phi(x) = 1, \quad \lim_{x \to -\infty} \Phi(x) = 0;$
2. $\Phi(0) = \frac{1}{2};$
3. $\Phi(-x) = 1 - \Phi(x)$, for all $x \in \mathbb{R}$.

Normal random variables

Now that we have seen the standard normal random variable, we can obtain any normal random variable by shifting and scaling a standard normal random variable. In particular, define

$$X = \sigma Z + \mu, \qquad \text{where } \sigma > 0.$$

Then

$$EX = \sigma EZ + \mu = \mu,$$

$$\text{Var}(X) = \sigma^2 \text{Var}(Z) = \sigma^2.$$

We say that X is a normal random variable with mean μ and variance σ^2. We write $X \sim N(\mu, \sigma^2)$.

If Z is a standard normal random variable and $X = \sigma Z + \mu$, then X is a normal random variable with mean μ and variance σ^2, i.e,

$$X \sim N(\mu, \sigma^2).$$

If X is a normal random variable with mean μ and variance σ^2, i.e, $X \sim N(\mu, \sigma^2)$, then

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\},$$

$$F_X(x) = P(X \leq x) = \Phi\left(\frac{x-\mu}{\sigma} \right),$$

$$P(a < X \leq b) = \Phi\left(\frac{b-\mu}{\sigma} \right) - \Phi\left(\frac{a-\mu}{\sigma} \right).$$

Example

Let $X \sim N(-5,4)$.

a.Find $P(X<0)$.

b.Find $P(-7<X<-3)$.

c.Find $P(X>-3 \mid X>-5)$

X is a normal random variable with $\mu = -5$ and $\sigma = \sqrt{4} = 2$, thus we have

a. Find $P(X < 0)$:

$$\begin{aligned} P(X < 0) &= F_X(0) \\ &= \Phi\left(\tfrac{0-(-5)}{2}\right) \\ &= \Phi(2.5) \approx 0.99 \end{aligned}$$

b. Find $P(-7 < X < -3)$:

$$\begin{aligned} P(-7 < X < -3) &= F_X(-3) - F_X(-7) \\ &= \Phi\left(\tfrac{(-3)-(-5)}{2}\right) - \Phi\left(\tfrac{(-7)-(-5)}{2}\right) \\ &= \Phi(1) - \Phi(-1) \\ &= 2\Phi(1) - 1 \qquad \left(\text{since } \Phi(-x) = 1 - \Phi(x)\right) \\ &\approx 0.68 \end{aligned}$$

c. Find $P(X > -3 | X > -5)$:

$$\begin{aligned} P(X > -3 | X > -5) &= \tfrac{P(X>-3, X>-5)}{P(X>-5)} \\ &= \tfrac{P(X>-3)}{P(X>-5)} \\ &= \frac{1-\Phi\left(\frac{(-3)-(-5)}{2}\right)}{1-\Phi\left(\frac{(-5)-(-5)}{2}\right)} \\ &= \tfrac{1-\Phi(1)}{1-\Phi(0)} \\ &\approx \tfrac{0.1587}{0.5} \approx 0.32 \end{aligned}$$

Theorem

If $X \sim N(\mu_X, \sigma_X^2)$, and $Y = aX + b$, where $a, b \in \mathbb{R}$, then $Y \sim N(\mu_Y, \sigma_Y^2)$ where

$$\mu_Y = a\mu_X + b, \quad \sigma_Y^2 = a^2\sigma_X^2.$$

Proof

We can write

$$X = \sigma_X Z + \mu_X \qquad \text{where } Z \sim N(0,1).$$

Thus,

$$\begin{aligned} Y &= aX + b \\ &= a(\sigma_X Z + \mu_X) + b \\ &= (a\sigma_X)Z + (a\mu_X + b). \end{aligned}$$

Therefore,

$$Y \sim N(a\mu_X + b, a^2\sigma_X^2).$$

Gamma Distribution

Gamma function: The gamma function, shown by $\Gamma(x)$, is an extension of the factorial function to real (and complex) numbers. Specifically, if $n \in \{1,2,3,\ldots\}$ then

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

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

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

Properties of the gamma function

For any positive real number α:

1. $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$;
2. $\int_0^\infty x^{\alpha-1} e^{-\lambda x} dx = \frac{\Gamma(\alpha)}{\lambda^\alpha}, \qquad \text{for } \lambda > 0$;
3. $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$;
4. $\Gamma(n) = (n-1)!, \text{ for } n = 1, 2, 3, \cdots$;
5. $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

Example

Answer the following questions:

1. Find $\Gamma(\frac{7}{2})$.
2. Find the value of the following integral:

$$I = \int_0^\infty x^6 e^{-5x} dx.$$

1. To find $\Gamma(\frac{7}{2})$, we can write

$$\begin{aligned}
\Gamma(\frac{7}{2}) &= \frac{5}{2} \cdot \Gamma(\frac{5}{2}) \qquad \text{(using Property 3)} \\
&= \frac{5}{2} \cdot \frac{3}{2} \cdot \Gamma(\frac{3}{2}) \qquad \text{(using Property 3)} \\
&= \frac{5}{2} \cdot \frac{3}{2} \cdot \frac{1}{2} \cdot \Gamma(\frac{1}{2}) \text{(using Property 3)} \\
&= \frac{5}{2} \cdot \frac{3}{2} \cdot \frac{1}{2} \cdot \sqrt{\pi} \qquad \text{(using Property 5)} \\
&= \frac{15}{8} \sqrt{\pi}.
\end{aligned}$$

2. Using Property 2 with $\alpha = 7$ and $\lambda = 5$, we obtain

$$
\begin{aligned}
I &= \int_0^\infty x^6 e^{-5x} dx \\
&= \frac{\Gamma(7)}{5^7} \\
&= \frac{6!}{5^7} \qquad \text{(using Property 4)} \\
&\approx 0.0092
\end{aligned}
$$

Gamma Distribution:

We now define the gamma distribution by providing its PDF:

A continuous random variable X is said to have a *gamma* distribution with parameters $\alpha > 0$ and $\lambda > 0$, shown as $X \sim Gamma(\alpha, \lambda)$, if its PDF is given by

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

If we let $\alpha = 1$, we obtain

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

Mixed Random Variables

- Here, we will discuss *mixed* random variables. These are random variables that are neither discrete nor continuous, but are a mixture of both. In particular, a mixed random variable has a continuous part and a discrete part. Thus, we can use our tools from previous chapters to analyze them.

In general, the CDF of a mixed random variable Y can be written as the sum of a continuous function and a staircase function:

$$F_Y(y) = C(y) + D(y).$$

We differentiate the continuous part of the CDF. In particular, let's define

$$c(y) = \frac{dC(y)}{dy}, \text{ wherever } C(y) \text{ is differentiable.}$$

Note that this is not a valid PDF as it does not integrate to one. Also, let $\{y_1, y_2, y_3, \dots\}$ be the set of jump points of $D(y)$, i.e., the points for which $P(Y = y_k) > 0$. We then have

$$\int_{-\infty}^{\infty} c(y)dy + \sum_{y_k} P(Y = y_k) = 1.$$

The expected value of Y can be obtained as

$$EY = \int_{-\infty}^{\infty} yc(y)dy + \sum_{y_k} y_k P(Y = y_k).$$

Multiple Random Variables

Joint Distributions and Independence

For three or more random variables, the joint PDF, joint PMF, and joint CDF are defined in a similar way to what we have already seen for the case of two random variables. Let X_1, X_2, $\cdots$, X_n be n discrete random variables. The joint PMF of X_1, X_2, $\cdots$, X_n is defined as

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

For n jointly continuous random variables X_1, X_2, $\cdots$, X_n, the joint PDF is defined to be the function $f_{X_1X_2\ldots X_n}(x_1,x_2,\ldots,x_n)$ such that the probability of any set $A \subset \mathbb{R}^n$ is given by the integral of the PDF over the set A. In particular, for a set $A \in \mathbb{R}^n$, we can write

$$P\Big((X_1,X_2,\cdots,X_n)\in A\Big) = \int\cdots\int_A \cdots \int f_{X_1X_2\cdots X_n}(x_1,x_2,\cdots,x_n)dx_1dx_2\cdots dx_n.$$

The marginal PDF of X_i can be obtained by integrating all other X_j's. For example,

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f_{X_1X_2\ldots X_n}(x_1,x_2,\ldots,x_n)dx_2\cdots dx_n.$$

The joint CDF of n random variables X_1, X_2,...,X_n is defined as

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

Independence

Independence: The idea of independence is exactly the same as what we have seen before. We restate it here in terms of the joint PMF, joint PDF, and joint CDF. Random variables X_1, X_2, ..., X_n are independent, if for all $(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$,

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

Equivalently, if X_1, X_2, ..., X_n are discrete, then they are independent if for all $(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$, we have

$$P_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = P_{X_1}(x_1)P_{X_2}(x_2)\cdots P_{X_n}(x_n).$$

If X_1, X_2, ..., X_n are continuous, then they are independent if for all $(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$, we have

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

If random variables X_1, X_2, ..., X_n are independent, then we have

$$E[X_1X_2\cdots X_n] = E[X_1]E[X_2]\cdots E[X_n].$$

Definition Random variables X_1, X_2, ..., X_n are said to be **independent and identically distributed (i.i.d.)** if they are *independent*, and they have the *same marginal distributions*:

$$F_{X_1}(x) = F_{X_2}(x) = \ldots = F_{X_n}(x), \text{ for all } x \in \mathbb{R}.$$

For example, if random variables X_1, X_2, ..., X_n are i.i.d., they will have the same means and variances, so we can write

$$\begin{aligned} E[X_1 X_2 \cdots X_n] &= E[X_1]E[X_2]\cdots E[X_n] && \text{(because the } X_i\text{'s are indepenednt)} \\ &= E[X_1]E[X_1]\cdots E[X_1] && \text{(because the } X_i\text{'s are identically distributed)} \\ &= E[X_1]^n. \end{aligned}$$

Random Vectors

When dealing with multiple random variables, it is sometimes useful to use vector and matrix notations. This makes the formulas more compact and lets us use facts from linear algebra. In this section, we briefly explore this avenue. The reader should be familiar with matrix algebra before reading this section. When we have n random variables X_1, X_2, ..., X_n we can put them in a (column) vector $\mathbf{X}$:

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \cdot \\ \cdot \\ \cdot \\ X_n \end{bmatrix}.$$

We call $\mathbf{X}$ a **random vector**. Here $\mathbf{X}$ is an n-dimensional vector because it consists of n random variables. In this book, we usually use bold capital letters such as $\mathbf{X}, \mathbf{Y}$ and $\mathbf{Z}$ to represent a random vector. To show a possible value of a random vector we usually use bold lowercase letters such as $\boldsymbol{x}$, $\mathbf{y}$ and $\mathbf{z}$. Thus, we can write the CDF of the random vector $\mathbf{X}$ as

$$\begin{aligned} F_{\mathbf{X}}(\boldsymbol{x}) &= F_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n) \\ &= P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n). \end{aligned}$$

If the X_i's are jointly continuous, the PDF of X can be written as

$$f_{\mathbf{X}}(\boldsymbol{x}) = f_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n).$$

Expectation

The **expected value vector** or the **mean vector** of the random vector **X** is defined as

$$E\mathbf{X} = \begin{bmatrix} EX_1 \\ EX_2 \\ \cdot \\ \cdot \\ \cdot \\ EX_n \end{bmatrix}.$$

Similarly, a **random matrix** is a matrix whose elements are random variables. In particular, we can have an m by n random matrix **M** as

$$\mathbf{M} = \begin{bmatrix} X_{11} & X_{12} & \dots & X_{1n} \\ X_{21} & X_{22} & \dots & X_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ X_{m1} & X_{m2} & \dots & X_{mn} \end{bmatrix}.$$

We sometimes write this as $\mathbf{M} = [X_{ij}]$, which means that X_{ij} is the element in the ith row and jth column of $\mathbf{M}$. The mean matrix of $\mathbf{M}$ is given by

$$E\mathbf{M} = \begin{bmatrix} EX_{11} & EX_{12} & \dots & EX_{1n} \\ EX_{21} & EX_{22} & \dots & EX_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ EX_{m1} & EX_{m2} & \dots & EX_{mn} \end{bmatrix}.$$

Linearity of expectation is also valid for random vectors and matrices. In particular, let $\mathbf{X}$ be an n-dimensional random vector and the random vector $\mathbf{Y}$ be defined as

$$\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{b},$$

where $\mathbf{A}$ is a fixed (non-random) m by n matrix and $\mathbf{b}$ is a fixed m-dimensional vector. Then we have

$$E\mathbf{Y} = \mathbf{A}E\mathbf{X} + \mathbf{b}.$$

Also, if $\mathbf{X_1}, \mathbf{X_2}, \cdots, \mathbf{X_k}$ are n-dimensional random vectors, then we have

$$\mathbf{E}[\mathbf{X_1} + \mathbf{X_2} + \cdots + \mathbf{X_k}] = \mathbf{E}\mathbf{X_1} + \mathbf{E}\mathbf{X_2} + \cdots + \mathbf{E}\mathbf{X_k}.$$

Correlation and Covariance Matrix

For a random vector $\mathbf{X}$, we define the **correlation matrix**, $\mathbf{R_X}$, as

$$\mathbf{R_X} = \mathbf{E}[\mathbf{X}\mathbf{X}^\mathbf{T}] = E\begin{bmatrix} X_1^2 & X_1X_2 & \dots & X_1X_n \\ X_2X_1 & X_2^2 & \dots & X_2X_n \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ X_nX_1 & X_nX_2 & \dots & X_n^2 \end{bmatrix} = \begin{bmatrix} EX_1^2 & E[X_1X_2] & \dots & E[X_1X_n] \\ EX_2X_1 & E[X_2^2] & \dots & E[X_2X_n] \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ E[X_nX_1] & E[X_nX_2] & \dots & E[X_n^2] \end{bmatrix}$$

where ^{T} shows matrix transposition.

The **covariance matrix**, C_X, is defined as

$$\mathbf{C_X} = \mathbf{E}[(\mathbf{X} - \mathbf{EX})(\mathbf{X} - \mathbf{EX})^{\mathrm{T}}]$$

$$= E \begin{bmatrix} (X_1 - EX_1)^2 & (X_1 - EX_1)(X_2 - EX_2) & \dots & (X_1 - EX_1)(X_n - EX_n) \\ (X_2 - EX_2)(X_1 - EX_1) & (X_2 - EX_2)^2 & \dots & (X_2 - EX_2)(X_n - EX_n) \\ \vdots & \vdots & \vdots & \vdots \\ (X_n - EX_n)(X_1 - EX_1) & (X_n - EX_n)(X_2 - EX_2) & \dots & (X_n - EX_n)^2 \end{bmatrix}$$

$$= \begin{bmatrix} \operatorname{Var}(X_1) & \operatorname{Cov}(X_1, X_2) & \dots & \operatorname{Cov}(X_1, X_n) \\ \operatorname{Cov}(X_2, X_1) & \operatorname{Var}(X_2) & \dots & \operatorname{Cov}(X_2, X_n) \\ \vdots & \vdots & \vdots & \vdots \\ \operatorname{Cov}(X_n, X_1) & \operatorname{Cov}(X_n X_2) & \dots & \operatorname{Var}(X_n) \end{bmatrix}.$$

The covariance matrix is a generalization of the variance of a random variable. Remember that for a random variable, we have $\operatorname{Var}(X) = EX^2 - (EX)^2$. The following example extends this formula to random vectors.

For a random vector $\mathbf{X}$, show

$$\mathbf{C_X} = \mathbf{R_X} - \mathbf{EXEX^T}.$$

Solution

We have

$$\begin{aligned}
\mathbf{C_X} &= \mathbf{E}[(\mathbf{X} - \mathbf{EX})(\mathbf{X} - \mathbf{EX})^\mathbf{T}] \\
&= \mathbf{E}[(\mathbf{X} - \mathbf{EX})(\mathbf{X^T} - \mathbf{EX^T})] \\
&= \mathbf{E}[\mathbf{XX^T}] - \mathbf{EXEX^T} - \mathbf{EXEX^T} + \mathbf{EXEX^T} \quad \text{(by linearity of expectation)} \\
&= \mathbf{R_X} - \mathbf{EXEX^T}.
\end{aligned}$$

Correlation matrix of $\mathbf{X}$:

$$\mathbf{R_X} = \mathbf{E}[\mathbf{X}\mathbf{X}^\mathbf{T}]$$

Covariance matrix of $\mathbf{X}$:

$$\mathbf{C_X} = \mathbf{E}[(\mathbf{X} - \mathbf{EX})(\mathbf{X} - \mathbf{EX})^\mathbf{T}] = \mathbf{R_X} - \mathbf{EXEX}^\mathbf{T}$$

Let $\mathbf{X}$ be an n-dimensional random vector and the random vector $\mathbf{Y}$ be defined as

$$\mathbf{Y} = \mathbf{AX} + \mathbf{b},$$

where $\mathbf{A}$ is a fixed m by n matrix and $\mathbf{b}$ is a fixed m-dimensional vector. Show that

$$\mathbf{C_Y} = \mathbf{AC_X A^T}.$$

Note that by linearity of expectation, we have

$$\mathbf{EY} = \mathbf{AEX} + \mathbf{b}.$$

By definition, we have

$$\begin{aligned}
\mathbf{C_Y} &= \mathbf{E}[(\mathbf{Y} - \mathbf{EY})(\mathbf{Y} - \mathbf{EY})^\mathbf{T}] \\
&= \mathbf{E}[(\mathbf{AX} + \mathbf{b} - \mathbf{AEX} - \mathbf{b})(\mathbf{AX} + \mathbf{b} - \mathbf{AEX} - \mathbf{b})^\mathbf{T}] \\
&= E[\mathbf{A}(\mathbf{X} - \mathbf{EX})(\mathbf{X} - \mathbf{EX})^\mathbf{T}\mathbf{A}^\mathbf{T}] \\
&= \mathbf{AE}[(\mathbf{X} - \mathbf{EX})(\mathbf{X} - \mathbf{EX})^\mathbf{T}]\mathbf{A}^\mathbf{T} && \text{(by linearity of expectation)} \\
&= \mathbf{AC_X A^T}.
\end{aligned}$$

Properties of the Covariance Matrix:

A ***covariance matrix*** $\Sigma \in \mathbb{R}^{n \times n}$ is simply a matrix such that there exists some random vector $X \in \mathbb{R}^n$ such that $\sigma_{ij} = \mathrm{Cov}(X_i, X_j)$ for all i and j.

Properties:

1. Σ is symmetric since $\mathrm{Cov}(X_i, X_j) = \mathrm{Cov}(X_j, X_i)$.
2. Σ is positive semi-definite (PSD):

$$\begin{aligned} u^T \Sigma u &= \sum_{i,j=1}^{n} u_i \Sigma_{ij} u_j \\ &= \sum_{i,j=1}^{n} \mathrm{Cov}(u_i X_i, u_j X_j) \\ &= \mathrm{Cov}\left(\sum_i u_i X_i, \sum_j u_j X_j \right) \geq 0. \end{aligned}$$

3. Because Σ is PSD, all of its eigenvalues are non-negative. (If Σ was positive definite, then its eigenvalues would be positive.)
4. Since Σ is real and symmetric, all of its eigenvalues are real, and there exists a real orthogonal matrix Q such that $D = Q^T \Sigma Q$ is a diagonal matrix. (The entries along the diagonal of D are Σ's eigenvalues.)
5. Since Σ's eigenvalues are all non-negative, Σ has a square root: $\Sigma^{1/2} = QD^{1/2}$. (If Σ is positive definite, then it has a negative square root as well: $\Sigma^{-1/2} = QD^{-1/2}$.)

Finally, if we have two random vectors, $\mathbf{X}$ and $\mathbf{Y}$, we can define the **cross correlation matrix** of $\mathbf{X}$ and $\mathbf{Y}$ as

$$\mathbf{R_{XY}} = \mathbf{E}[\mathbf{X}\mathbf{Y}^\mathbf{T}].$$

Also, the **cross covariance matrix** of $\mathbf{X}$ and $\mathbf{Y}$ is

$$\mathbf{C_{XY}} = \mathbf{E}[(\mathbf{X} - \mathbf{EX})(\mathbf{Y} - \mathbf{EY})^\mathbf{T}].$$

Functions of Random Vectors: The Method of Transformations

A function of a random vector is a random vector. Thus, the methods that we discussed regarding functions of two random variables can be used to find distributions of functions of random vectors. For example, we can state a more general form of Theorem method of transformations . Let us first explain the method and then see some examples on how to use it. Let $\mathbf{X}$ be an n-dimensional random vector with joint PDF $f_{\mathbf{X}}(\mathbf{x})$. Let $G : \mathbb{R}^n \mapsto \mathbb{R}^n$ be a continuous and invertible function with continuous partial derivatives and let $H = G^{-1}$. Suppose that the random vector $\mathbf{Y}$ is given by $\mathbf{Y} = G(\mathbf{X})$ and thus $\mathbf{X} = G^{-1}(\mathbf{Y}) = H(\mathbf{Y})$. That is,

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \cdot \\ \cdot \\ \cdot \\ X_n \end{bmatrix} = \begin{bmatrix} H_1(Y_1, Y_2, \ldots, Y_n) \\ H_2(Y_1, Y_2, \ldots, Y_n) \\ \cdot \\ \cdot \\ \cdot \\ H_n(Y_1, Y_2, \ldots, Y_n) \end{bmatrix}.$$

Then, the PDF of Y, $f_{Y_1, Y_2, \ldots, Y_n}(y_1, y_2, \ldots, y_n)$, is given by

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}\big(H(\mathbf{y})\big)|J|$$

where J is the Jacobian of H defined by

$$J = \det \begin{bmatrix} \frac{\partial H_1}{\partial y_1} & \frac{\partial H_1}{\partial y_2} & \cdots & \frac{\partial H_1}{\partial y_n} \\ \frac{\partial H_2}{\partial y_1} & \frac{\partial H_2}{\partial y_2} & \cdots & \frac{\partial H_2}{\partial y_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial H_n}{\partial y_1} & \frac{\partial H_n}{\partial y_2} & \cdots & \frac{\partial H_n}{\partial y_n} \end{bmatrix}^{-1}$$

and evaluated at $(y_1, y_2, \ldots, y_n)$.

Normal (Gaussian) Random Vectors:

Random variables $X_1, X_2, \ldots, X_n$ are said to be **jointly normal** if, for all $a_1, a_2, \ldots, a_n \in \mathbb{R}$, the random variable

$$a_1X_1 + a_2X_2 + \ldots + a_nX_n$$

is a normal random variable.

A random vector

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \cdot \\ \cdot \\ \cdot \\ X_n \end{bmatrix}$$

is said to be **normal** or **Gaussian** if the random variables $X_1, X_2, \ldots, X_n$ are jointly normal.

For a standard normal random vector $\mathbf{Z}$, where Z_i's are i.i.d. and $Z_i \sim N(0,1)$, the PDF is given by

$$f_{\mathbf{Z}}(\mathbf{z}) = \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\left\{-\frac{1}{2}\mathbf{z}^T\mathbf{z}\right\}.$$

For a normal random vector $\mathbf{X}$ with mean $\mathbf{m}$ and covariance matrix $\mathbf{C}$, the PDF is given by

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{\det \mathbf{C}}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^T\mathbf{C}^{-1}(\mathbf{x}-\mathbf{m})\right\}$$

Central Limit Theorem

The Central Limit Theorem (CLT)

Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with expected value $EX_i = \mu < \infty$ and variance $0 < \mathrm{Var}(X_i) = \sigma^2 < \infty$. Then, the random variable

$$Z_n = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{X_1 + X_2 + \ldots + X_n - n\mu}{\sqrt{n}\sigma}$$

converges in distribution to the standard normal random variable as n goes to infinity, that is

$$\lim_{n \to \infty} P(Z_n \leq x) = \Phi(x), \qquad \text{for all } x \in \mathbb{R},$$

where $\Phi(x)$ is the standard normal CDF.

How to Apply The Central Limit Theorem (CLT)

Here are the steps that we need in order to apply the CLT:

1. Write the random variable of interest, Y, as the sum of n i.i.d. random variable X_i's:

$$Y = X_1 + X_2 + \ldots + X_n.$$

2. Find EY and $\mathbf{Var}(Y)$ by noting that

$$EY = n\mu, \qquad \mathbf{Var}(Y) = n\sigma^2,$$

where $\mu = EX_i$ and $\sigma^2 = \mathbf{Var}(X_i)$.

3. According to the CLT, conclude that $\frac{Y - EY}{\sqrt{\mathbf{Var}(Y)}} = \frac{Y - n\mu}{\sqrt{n}\sigma}$ is approximately standard normal; thus, to find $P(y_1 \leq Y \leq y_2)$, we can write

$$\begin{aligned} P(y_1 \leq Y \leq y_2) &= P\left(\frac{y_1 - n\mu}{\sqrt{n}\sigma} \leq \frac{Y - n\mu}{\sqrt{n}\sigma} \leq \frac{y_2 - n\mu}{\sqrt{n}\sigma}\right) \\ &\approx \Phi\left(\frac{y_2 - n\mu}{\sqrt{n}\sigma}\right) - \Phi\left(\frac{y_1 - n\mu}{\sqrt{n}\sigma}\right). \end{aligned}$$

Introduction to Random Processes

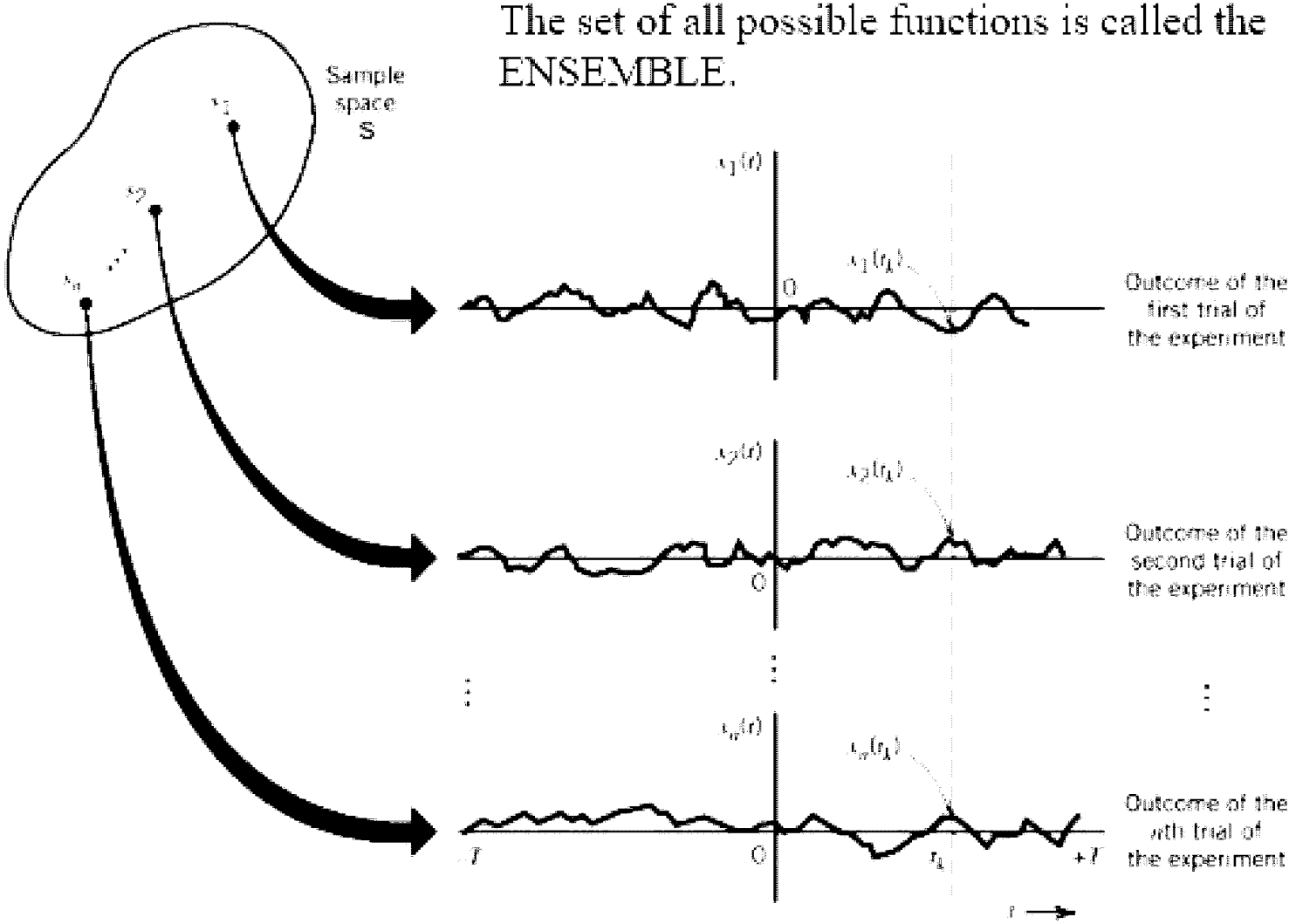
- Description of Random Processes
- Stationarity and Ergodicty
- Autocorrelation of Random Processes
- Properties of Autocorrelation

Random Processes

- A **RANDOM VARIABLE** X, is a rule for assigning to every outcome, ω, of an experiment a number $X(\omega)$.
  - Note: X denotes a random variable and $X(\omega)$ denotes a particular value.
- A **RANDOM PROCESS** $X(t)$ is a rule for assigning to every ω, a function $X(t,\omega)$.
  - Note: for notational simplicity we often omit the dependence on ω.

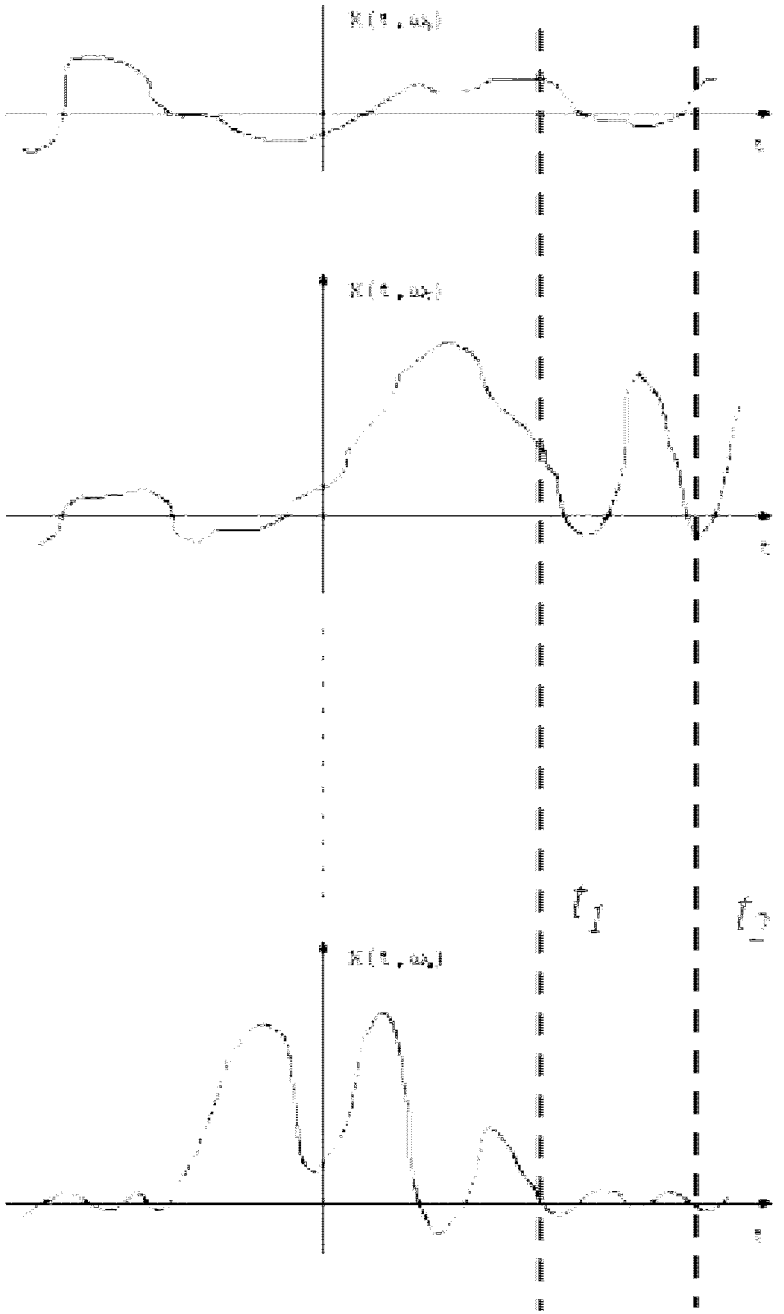
Ensemble of Sample Functions

The set of all possible functions is called the ENSEMBLE.



Random Processes

- A general Random or Stochastic Process can be described as:
  - Collection of time functions (signals) corresponding to various outcomes of random experiments.
  - Collection of random variables observed at different times.
- Examples of random processes in communications:
  - Channel noise.
  - Information generated by a source.
  - Interference.

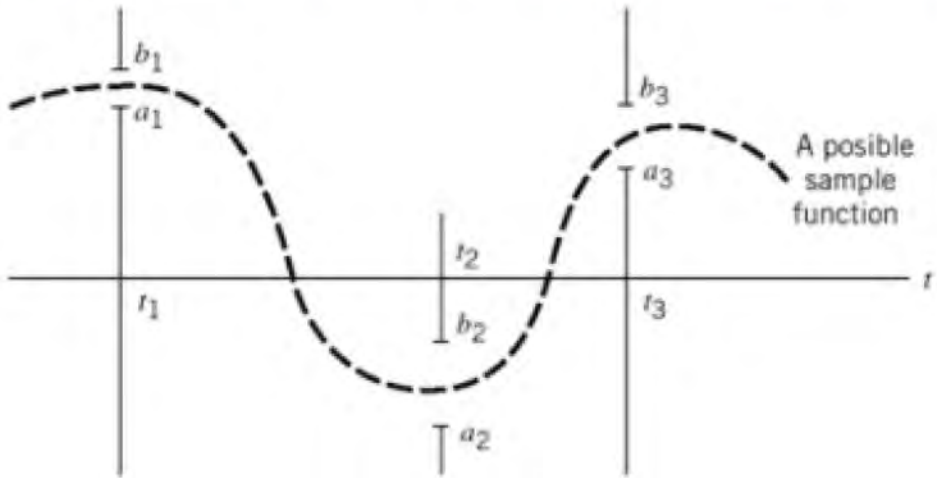


Collection of Time Functions

- Consider the time-varying function representing a random process where ω_i represents an outcome of a random event.
- Example:
  - a box has infinitely many resistors ($i=1,2,\ldots$) of same resistance R.
  - Let ω_i be event that the ith resistor has been picked up from the box
  - Let $v(t, \omega_i)$ represent the voltage of the thermal noise measured on this resistor.

Collection of Random Variables

- For a particular time $t=t_o$ the value $x(t_o,\omega_i)$ is a random variable.
- To describe a random process we can use collection of random variables $\{x(t_o,\omega_1)\ ,\ x(t_o,\omega_2)\ ,\ x(t_o,\omega_3)\ ,\ \ldots\ \}$.
- Type: a random processes can be either discrete-time or continuous-time.
- Probability of obtaining a sample function of a RP that passes through the following set of windows. Probability of a joint event.



Description of Random Processes

- **Analytical description:** $X(t) = f(t, \omega)$ where ω is an outcome of a random event.
- **Statistical description:** For any integer N and any choice of $(t_1, t_2, \ldots, t_N)$ the joint pdf of $\{X(t_1), X(t_2), \ldots, X(t_N)\}$ is known. To describe the random process completely the PDF $f(\mathbf{x})$ is required.

$$x_1 = x(t_1), \quad \mathbf{x} = [x_1, x_2, \ldots, x_N]$$

$$f(\mathbf{x}) = f\{x(t_1), x(t_2), \ldots, x(t_N)\}$$

Example: Analytical Description

- Let $X(t) = A\cos(2\pi f_0 t + \theta)$ where θ is a random variable uniformly distributed on $[0,2\pi]$.
- Complete statistical description of $X(t_o)$ is:
  - Introduce $Y = 2\pi f_0 t_0 + \theta$
  - Then, we need to transform from y to x:

$$p_X(x)dx = p_Y(y_1)dy + p_Y(y_2)dy$$

  - We need both y_1 and y_2 because for a given x the equation $x=A\cos(y)$ has two solutions in $[0,2\pi]$.

Analytical (continued)

- Note x and y are actual values of the random variables X and Y.
- Since $\left|\frac{dx}{dt}\right| = |A\sin y| = \left|\sqrt{A^2 - x^2}\right|$

  and p_Y is uniform in $[2\pi f_0 t_0,\ 2\pi f_0 t_0 + 2\pi]$, we get

$$p_X(x) = \begin{cases} \dfrac{1}{\pi\sqrt{A^2 - x^2}} & -A < x < A \\ 0 & \text{Elsewhere} \end{cases}$$

- Using the analytical description of $X(t)$, we obtained its statistical description at any time t.

Example: Statistical Description

- Suppose a random process x(t) has the property that for any N and (t_0,t_1, . . .,t_N) the joint density function of $\{x(t_i)\}$ is a jointly distributed Gaussian vector with zero mean and covariance

$$\sigma_{ij} = \sigma^2 \min(t_i, t_j)$$

- This gives complete statistical description of the random process x(t).

Stationarity

- Definition: A random process is STATIONARY to the order N if for any $t_1, t_2, \ldots, t_N$,

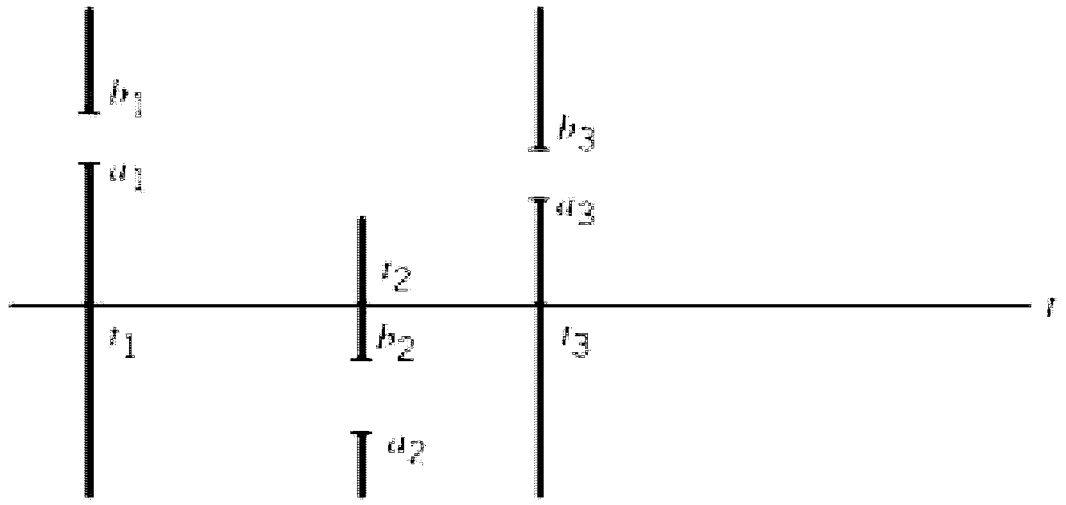
$$f_x\{x(t_1), x(t_2), \ldots x(t_N)\} = f_x\{x(t_1+t_0), x(t_2+t_0), \ldots, x(t_N+t_0)\}$$

- This means that the process behaves similarly (follows the same PDF) regardless of when you measure it.

- A random process is said to be STRICTLY STATIONARY if it is stationary to the order of $N \rightarrow \infty$.

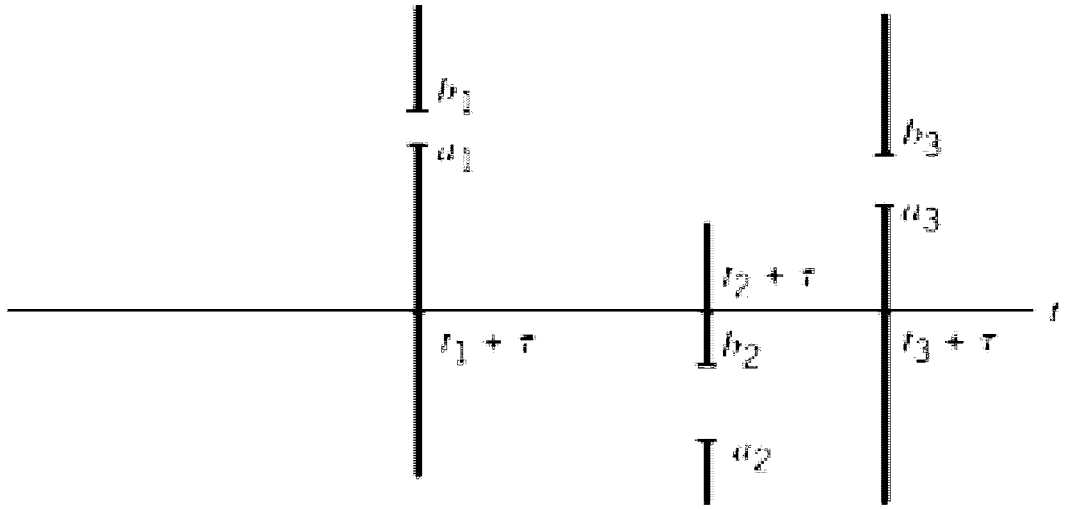
- Is the random process from the coin tossing experiment stationary?

Illustration of Stationarity



(a)

Time functions pass through the corresponding windows at different times with the same probability.



Example of First-Order Stationarity

RANDOM PROCESS is $x(t) = A\sin(\omega_0 t + \theta_0)$

- Assume that A and ω_0 are constants; θ_0 is a uniformly distributed RV from $[-\pi,\pi)$; t is time.
- From last lecture, recall that the PDF of *x(t):*

$$f_x(x) = \begin{cases} \dfrac{1}{\pi\sqrt{A^2 - x^2}} & |x| \le A \\ 0 & x \text{ Elsewhere} \end{cases}$$

- *Note*: there is NO dependence on time, the PDF is not a function of *t.*
- The RP is STATIONARY .

Non-Stationary Example

RANDOM PROCESS is $x(t) = A\sin(\omega_0 t + \theta_0)$

- Now assume that A, θ_0 and ω_0 are constants; t is time.
- Value of x(t) is <u>always</u> known for any time with a probability of 1. Thus the first order PDF of x(t) is

$$f(x) = \delta\left(x - A\sin\left(\omega_0 t + \theta_0\right)\right)$$

- *Note*: The PDF depends on time, so it is NONSTATIONARY .

Ergodic Processes

- Definition: A random process is *ERGODIC* if all time averages of any sample function are equal to the corresponding ensemble averages (expectations)
- Example, for ergodic processes, can use ensemble statistics to compute DC values and RMS values

$$x_{DC} = \langle x(t) \rangle = \overline{[x(t)]} = m_x$$

$$\langle x(t) \rangle = \lim_{T \to \infty} \frac{1}{T} \int_0^T [x(t)]dt \qquad \text{Time average}$$

$$\overline{[x(t)]} = \int_{-\infty}^{\infty} [x] f(x)dx = m_x \qquad \text{Ensmble average}$$

$$x_{RMS} = \sqrt{\langle x^2(t) \rangle} = \sqrt{\overline{x^2}} = \sqrt{\sigma^2 + m_x^2}$$

- Ergodic processes are always stationary; Stationary processes are not necessarily ergodic

$$\text{Ergodic} \Rightarrow \text{Stationary}$$

Example: Ergodic Process

RANDOM PROCESS is $x(t) = A\sin(\omega_0 t + \theta_0)$

- A and ω_0 are constants; θ_0 is a uniformly distributed RV from $[-\pi,\pi)$; t is time.
- Mean (Ensemble statistics)

$$m_x = \bar{x} = \int_{-\infty}^{\infty} x(\theta) f_\theta(\theta) d\theta = \int_{-\pi}^{\pi} A\sin(\omega_0 t + \theta) \frac{1}{2\pi} d\theta = 0$$

- Variance

$$\sigma_x^{\,2} = \int_{-\pi}^{\pi} A^2 \sin^2(\omega_0 t + \theta) \frac{1}{2\pi} d\theta = \frac{A^2}{2}$$

Example: Ergodic Process

- Mean (Time Average) T is large

$$\langle x(t) \rangle = \lim_{T \to \infty} \frac{1}{T} \int_0^T A \sin(\omega_0 t + \theta)\, dt = 0$$

- Variance

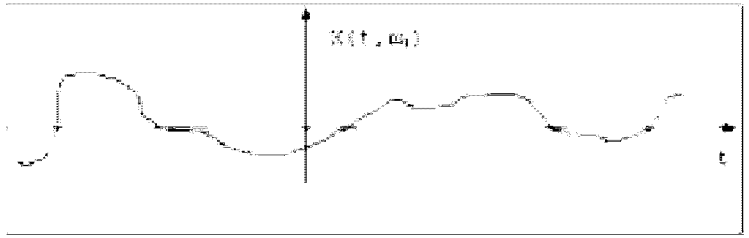
$$\langle x^2(t) \rangle = \lim_{T \to \infty} \frac{1}{T} \int_0^T A^2 \sin^2(\omega_0 t + \theta)\, dt = \frac{A^2}{2}$$

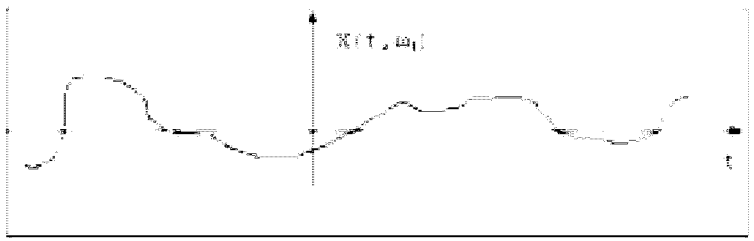
- The ensemble and time averages are the same, so the process is ERGODIC

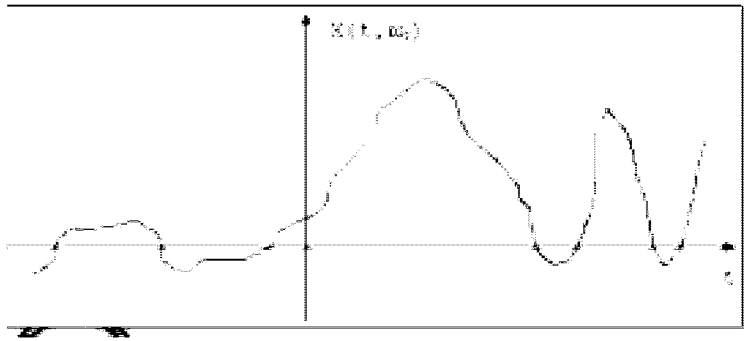
Autocorrelation of Random Process

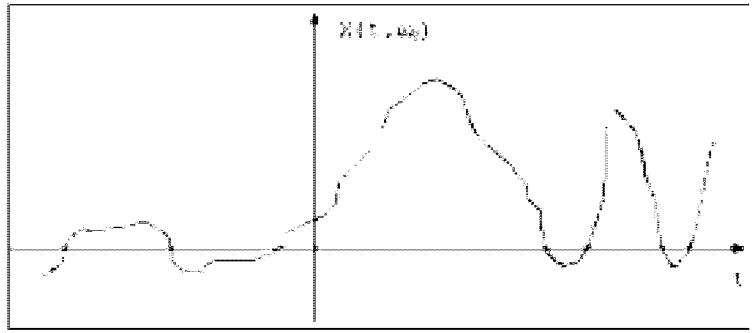
- The Autocorrelation function of a real random process x(t) at two times is:

$$R_x(t_1,t_2) = \overline{x(t_1)x(t_2)} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1 x_2 f_x(x_1,x_2)\,dx_1 dx_2$$









Wide-sense Stationary

- A random process that is stationary to order 2 or greater is *Wide-Sense Stationary*.
- A random process is *Wide-Sense Stationary* if:

$$\overline{x(t)} = \text{constant}$$

$$R_x(t_1, t_2) = R_x(\tau)$$

- Usually, $t_1 = t$ and $t_2 = t + \tau$ so that $t_2 - t_1 = \tau$.
- Wide-sense stationary process does not DRIFT with time.
- Autocorrelation depends only on the time gap but not where the time difference is.
- Autocorrelation gives idea about the frequency response of the RP.

Autocorrelation Function of RP

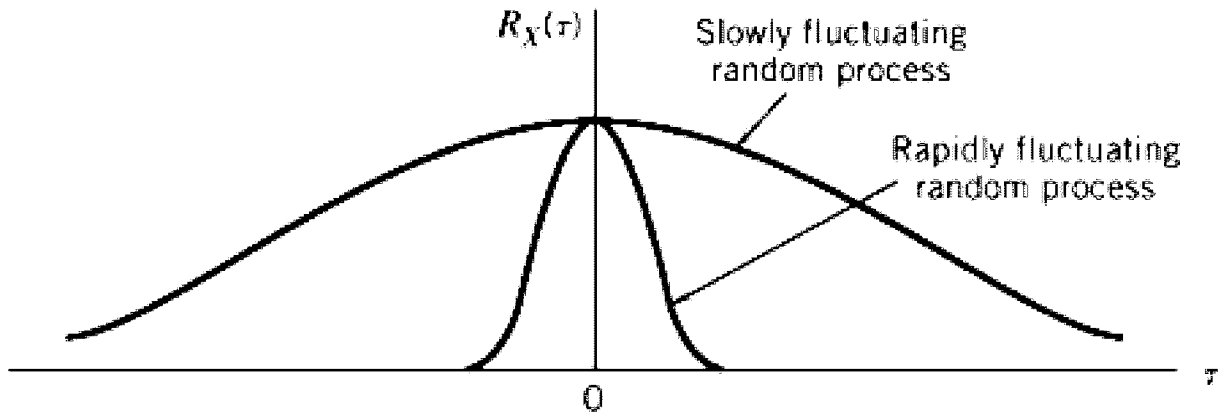
- Properties of the autocorrelation function of wide-sense stationary processes

$$R_x(0) = \overline{x^2(t)} = \text{Second Moment}$$

$$R_x(\tau) = R_x(-\tau), \quad \text{Symmetric}$$

$$R_x(0) \geq |R_x(\tau)|, \quad \text{Maximum value at 0}$$

Autocorrelation of slowly and rapidly fluctuating random processes.



Cross Correlations of RP

- Cross Correlation of two RP $x(t)$ and $y(t)$ is defined similarly as:

$$R_{xy}(t_1, t_2) = \overline{x(t_1)\, y(t_2)} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_1 y_2 f_{xy}(x_1, y_2)\, dx_1 dy_2$$

- If $x(t)$ and $y(t)$ are Jointly Stationary processes,

$$R_{xy}(t_1, t_2) = R_{xy}(t_2 - t_1) = R_{xy}(\tau) \qquad \tau = t_2 - t_1$$

- If the RP's are jointly ERGODIC,

$$R_{xy}(\tau) = \overline{x(t)\, y(t+\tau)} = \langle x(t)\, y(t+\tau) \rangle$$

Cross Correlation Properties of Jointly Stationary RP's

- Some properties of cross-correlation functions are

$$R_{xy}(\tau) = R_{xy}(-\tau)$$

$$\left|R_{xy}(\tau)\right| \le \sqrt{R_x(0)R_y(0)}$$

$$\left|R_{xy}(\tau)\right| \le \frac{1}{2}\left[R_x(0)+R_y(0)\right]$$

- Uncorrelated:

$$R_{xy}(\tau) = \overline{x(t)y(t+\tau)} = \bar{x}\,\bar{y}$$

- Orthogonal:

$$R_{xy}(\tau) = 0$$

- Independent: if $x(t_1)$ and $y(t_2)$ are independent (joint distribution is product of individual distributions)