

Probability and Random Variables
EE230 Section 04
2023 Spring

Barış Nakiboğlu*

April 18, 2023

Contents

1	A Review Of Set Theory	3
1.1	Representations of Sets	3
1.2	Set Operations	4
1.3	Algebra of Sets: Properties of Set Operations	5
2	Probability Spaces (Probability Models)	9
2.1	The Set of All Outcomes: The Sample Space Ω	9
2.2	The Set of All Events: The σ -algebra \mathcal{F}	10
2.3	Probability Law and Probability Axioms	12
2.3.1	Probability Axioms	12
2.3.2	Properties of Probability Laws	12
2.4	Discrete Probability Models	14
2.5	Continuous Probability Models	16
3	Conditional Probability	19
3.1	Conditional Probability of An Event Given Another Event	19
3.2	Conditional Probability Law	20
3.3	Multiplication Rule (Chain Rule) and Modeling	21
3.4	Total Probability Theorem	23
3.5	Bayes' Rule	24
4	Independence	27
4.1	Independence Of Two Events	27
4.2	Conditional Independence	28
4.3	Independence of A Collection of Events	29
4.4	Independent Trials	31
5	Counting	33

*These lecture notes are prepared as a companion to the textbook [BT08] for exclusive private use of students in Section 04 during 2023 Spring term. Most of the examples are from [BT08].

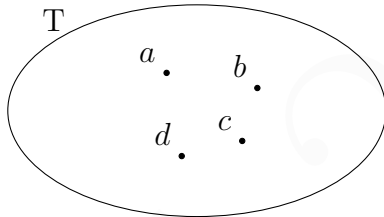
6	Discrete Random Variables	37
6.1	Probability Mass Function	38
6.2	Functions of A Random Variable	43
6.3	Expected Value and Variance of a Random Variable	44
6.4	Joint Probability Mass Function	49
6.5	Functions of Multiple Random Variables	51
6.6	Conditioning	53
6.6.1	Conditional PMFs Given an Event	53
6.6.2	Conditional PMFs Given A Discrete Random Variable	55
6.7	Conditional Expectation and Iterated Expectations	58
6.7.1	Conditional Variance and Law of Total Variance	59
6.8	Independence	62
6.8.1	Independence of a Random Variable From an Event	62
6.8.2	Independence of Two Random Variables	63
6.8.3	Independence of Several Random Variables	65

1 A Review Of Set Theory

Definition 1.1. A *set* is a collection of distinct elements.

$$\begin{aligned}
 x \in S & \iff x \text{ is an element of } S, \text{ i.e., } x \text{ belongs to } S. \\
 x \notin S & \iff x \text{ is not an element of } S, \text{ i.e., } x \text{ does not belong to } S.
 \end{aligned}$$

1.1 Representations of Sets



Venn Diagram

$$\begin{aligned}
 T &= \{a, b, c, d\} \\
 \mathbb{Z}_+ &= \{1, 2, \dots\} \\
 \mathbb{Z} &: \text{The set of all integers.} \\
 \mathbb{Q} &: \text{The set of all rational numbers integers.} \\
 \mathbb{R} &: \text{The set of all real numbers.} \\
 A &= \{x \in \mathbb{R} : \sin(x) \geq \frac{1}{2}\}
 \end{aligned}$$

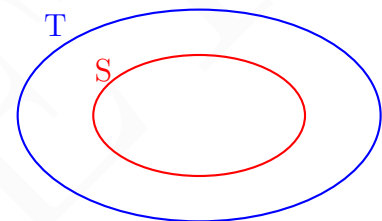
Definition 1.2. The *cardinality* (size) of a set S is the number of elements and denoted by $|S|$.
 A set S is *finite* iff $|S|$ is an integer, i.e., is finite. Sets that are not finite, i.e., sets with infinitely many elements, are called *infinite* sets.

- T is a finite set.
- $\mathbb{Z}, \mathbb{Q}, \mathbb{R}, A$ are infinite sets.

Definition 1.3. The *empty set* \emptyset is the set that does not have any elements. It is also called the *null set* and denoted by $\{\}$.

Definition 1.4. The *universal set* Ω is the set that contains all objects of interest, i.e., all possible elements, in a particular context.

Definition 1.5. If every element of a set S is an element of a set T , then S is a subset of T and denoted by $S \subset T$ or $T \supset S$.



Note that any set S is a subset of itself, i.e., $S \subset S$.

Definition 1.6. A set S is equal to a set T if and only if every element of S is an element of T and every element of T is an element S , i.e.. “ $S \subset T$ and $T \subset S$ ” $\iff S = T$

Definition 1.7. The *power set* 2^T of set is the set of all subsets of the set T :

$$2^T := \{S : S \subset T\} \tag{1.1}$$

Power set 2^T is also denoted as $\wp(T)$.

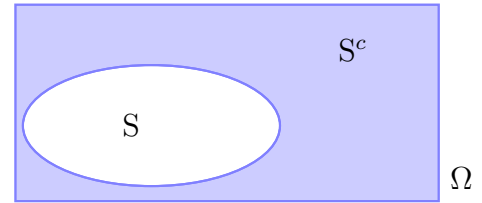
For a finite set T

$$|2^T| = 2^{|T|}. \tag{1.2}$$

1.2 Set Operations

i) The complement of a set with respect to the universal set.

$$S^c := \{x \in \Omega : x \notin S\}.$$

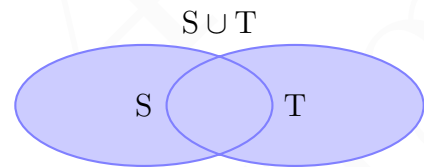


ii) The union of sets

$$S \cup T := \{x : x \in S \text{ or } x \in T\}$$

$$\bigcup_{j=1}^k S_j = \{x : x \in S_j \text{ for some } j \in \{1, 2, \dots, k\}\}$$

$$\bigcup_{j \in J} S_j = \{x : x \in S_j \text{ for some } j \in J\}$$



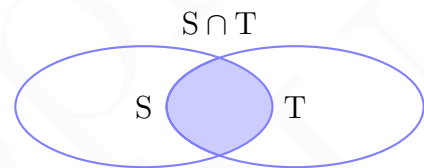
Index set J can be finite or infinite.

iii) The intersection of sets

$$S \cap T := \{x : x \in S \text{ and } x \in T\}$$

$$\bigcap_{j=1}^k S_j = \{x : x \in S_j \text{ for all } j \in \{1, 2, \dots, k\}\}$$

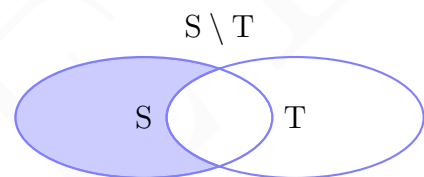
$$\bigcap_{j \in J} S_j = \{x : x \in S_j \text{ for all } j \in J\}$$



Index set J can be finite or infinite.

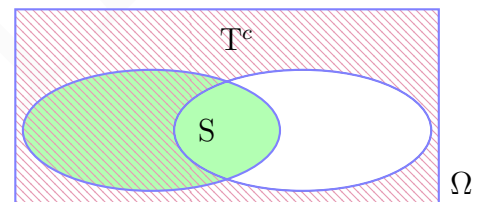
iv) The difference of two sets

$$S \setminus T := \{x : x \in S \text{ and } x \notin T\}$$



Often there exists a universal set Ω and we can write the difference using the complement as follows,

$$\begin{aligned} S \setminus T &= \{x \in \Omega : x \in S \text{ and } x \notin T\} \\ &= \{x \in \Omega : x \in S \text{ and } x \in T^c\} \\ &= S \cap T^c \end{aligned}$$



1.3 Algebra of Sets: Properties of Set Operations

1. Commutative property:

$$\begin{aligned}S \cup T &= T \cup S \\S \cap T &= T \cap S\end{aligned}$$

2. Associative property:

$$\begin{aligned}S \cup (T \cup Y) &= (S \cup T) \cup Y \\S \cap (T \cap Y) &= (S \cap T) \cap Y\end{aligned}$$

3. Distributive property:

$$\begin{aligned}S \cap (T \cup Y) &= (S \cap T) \cup (S \cap Y) \\S \cup (T \cap Y) &= (S \cup T) \cap (S \cup Y)\end{aligned}$$

4. Identity elements

$$\begin{aligned}S \cup \emptyset &= S \\S \cap \Omega &= S\end{aligned}$$

5. Complements

$$\begin{aligned}S \cup S^c &= \Omega \\S \cap S^c &= \emptyset\end{aligned}$$

Remark. These five properties can be confirmed using the definitions and the assertions we have in the following can be deduced from them. But for our purposes this distinction will be of minor importance.

6. Idempotent laws:

$$\begin{aligned}S \cup S &= S \\S \cap S &= S\end{aligned}$$

Proof of Idempotent law $S \cup S = S$.

$$\begin{aligned}S \cup S &= (S \cup S) \cap \Omega && \text{because } \Omega \text{ is the identity element for the intersection,} \\ &= (S \cup S) \cap (S \cup S^c) && \text{because } \Omega = S \cup S^c, \\ &= S \cup (S \cap S^c) && \text{because of the distributivity of union over intersection,} \\ &= S \cup \emptyset && \text{because } \emptyset = S \cap S^c, \\ &= S && \text{because } \emptyset \text{ is the identity element for the union.}\end{aligned}$$

□

7. Domination laws:

$$\begin{aligned}S \cup \Omega &= \Omega \\S \cap \emptyset &= \emptyset\end{aligned}$$

8. Absorption laws:

$$S \cup (S \cap T) = S$$

$$S \cap (S \cup T) = S$$

9. Involution law (Double complement law)

$$(S^c)^c = S$$

10. Complement laws for the universe set and the empty set:

$$\Omega^c = \emptyset$$

$$\emptyset^c = \Omega$$

Proof of $\Omega^c = \emptyset$.

$$\begin{aligned} \emptyset &= \Omega^c \cap \Omega \\ &= \Omega^c \end{aligned}$$

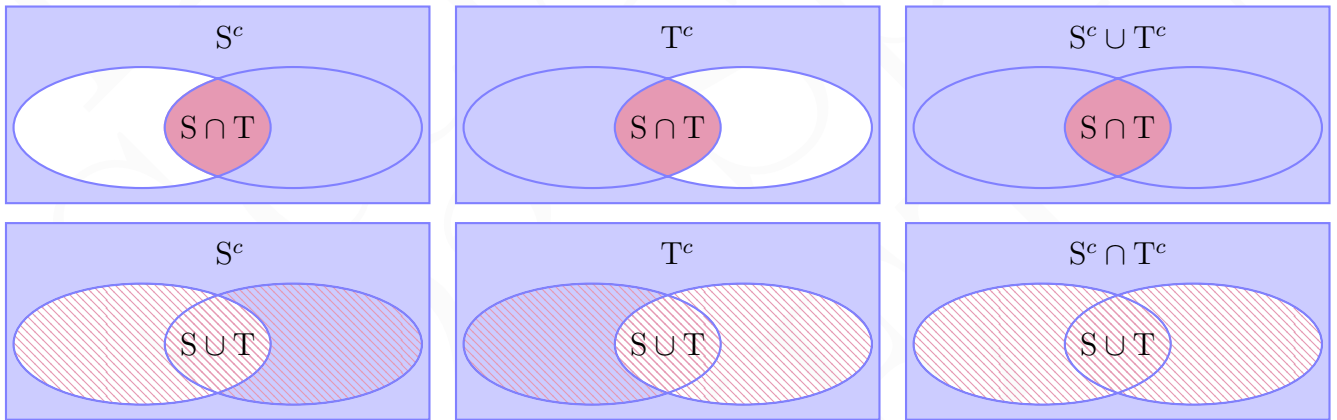
$$\begin{aligned} &\text{because } S^c \cap S = \emptyset, \\ &\text{because } S \cap \Omega = S. \end{aligned}$$

□

11. De Morgan's Law:

$$(S \cap T)^c = S^c \cup T^c$$

$$(S \cup T)^c = S^c \cap T^c$$



Proof of De Morgan's Laws $(S \cap T)^c = S^c \cup T^c$.

$$\begin{aligned} (S \cap T)^c &= \{x \in \Omega : x \notin (S \cap T)\} \\ &= \{x \in \Omega : \text{either } x \notin S \text{ or } x \notin T\} \\ &= \{x \in \Omega : \text{either } x \in S^c \text{ or } x \in T^c\} \\ &= S^c \cup T^c \end{aligned}$$

This proof uses logic, a proof relying only on the first five properties can be found. □

Remark. De Morgan's Law has the following more general form

$$\begin{aligned} \left(\bigcap_{j \in J} S_j \right)^c &= \bigcup_{j \in J} S_j^c \\ \left(\bigcup_{j \in J} S_j \right)^c &= \bigcap_{j \in J} S_j^c \end{aligned}$$

Proof of De Morgan's Laws $\left(\bigcup_{j \in J} S_j\right)^c = \bigcap_{j \in J} S_j^c$.

$$\begin{aligned} \left(\bigcup_{j \in J} S_j\right)^c &= \{x \in \Omega : \nexists j \in J \text{ such that } x \in S_j\} \\ &= \{x \in \Omega : x \notin S_j \forall j \in J\} \\ &= \{x \in \Omega : x \in S_j^c \forall j \in J\} \\ &= \bigcap_{j \in J} S_j^c \end{aligned}$$

□

Definition 1.8. The *cartesian product* of a set S and a set T is the set of all ordered pairs (s, t) such that $s \in S$ and $t \in T$:

$$S \times T = \{(s, t) : s \in S \text{ and } t \in T\}.$$

If S and T are finite sets then

$$|S \times T| = |S| \cdot |T|.$$

Definition 1.9. A set S is *countable* iff there exists a 1-to-1 correspondence between S and \mathbb{Z}_+ , i.e., can be written as a list. Sets that are not countable, i.e. sets that cannot be written as a list, are called *uncountable* sets.

- Any finite set is also a countable set.
- Any uncountable set is an infinite set. (uncountably infinite)
- Infinite and countable set are also called countably infinite sets.

a) If S and T are countable then $S \cup T$ is countable.

$$S \cup T = \{s_1, t_1, s_2, t_2, \dots\}$$

b) If S_j is countable for all $j \in J$ and J is countable then $\bigcup_{j \in J} S_j$ is countable.

	1	2	3	...
S_1	$s_{1,1}$	$s_{1,2}$	$s_{1,3}$...
S_2	$s_{2,1}$	$s_{2,2}$	$s_{2,3}$...
S_3	$s_{3,1}$	$s_{3,2}$	$s_{3,3}$...
\vdots	\vdots	\vdots	\vdots	\ddots

c) If S and T are countable then $S \times T$ is countable. Hence \mathbb{Q} is countable.

	t_1	t_2	...
s_1	(s_1, t_1)	(s_1, t_2)	...
s_2	(s_2, t_1)	(s_2, t_2)	...
\vdots	\vdots	\vdots	\ddots

d) If S_j is countable for all $j \in \{1, \dots, \kappa\}$ then $S_1 \times \dots \times S_\kappa$ is countable.

e) \mathbb{R} and any open interval (a, b) are uncountable sets.

The uncountability of a set of real numbers is proved via Cantor's diagonalisation argument.

If you are interested a short video explaining it can be found here:

<https://www.youtube.com/watch?v=YIZd23zGV3M>

Definition 1.10. Two sets are *disjoint* (*mutually exclusive*) iff their intersection is the empty set, i.e.,

$$\text{"S and T are disjoint"} \iff S \cap T = \emptyset.$$

A collection of sets $\{S_j\}_{j \in J}$ is disjoint iff $S_i \cap S_j = \emptyset$ for all $i \neq j$ in J .

Definition 1.11. A collection of sets $\{S_j\}_{j \in J}$ is said to be a *partition* of set T iff sets are disjoint and their union is equal to T , i.e. iff

$$\begin{aligned} S_i \cap S_j &= \emptyset & \forall i, j \in J : i \neq j \\ \bigcup_{j \in J} S_j &= T \end{aligned}$$

2 Probability Spaces (Probability Models)

There are many events/systems that are very hard or impossible to analyze via deterministic models. The probability theory provides us a language to describe and analyze these systems as random experiments. A probability space $(\Omega, \mathcal{F}, \mathbf{P}(\cdot))$ is a mathematical model that describes random experiment.

The samples space Ω is the set of all possible outcomes of the random experiment; hence the elements of Ω are the outcomes of the random experiment.

The σ -algebra \mathcal{F} is a collection of subsets of Ω for which probabilistic model is required to determine the probabilities. The elements of \mathcal{F} are called events. Thus \mathcal{F} is also called set of all events.

The probability law $\mathbf{P}(\cdot)$ is a function determining the probabilities of all of the events, i.e., all of the elements of \mathcal{F} .

This high level description of the sample space Ω , the σ -algebra \mathcal{F} , and the probability law $\mathbf{P}(\cdot)$ explains the role they play in a probability space. However, in order to play those roles in a probability space, the set of all events \mathcal{F} should be a σ -algebra and the probability law $\mathbf{P}(\cdot)$ should satisfy the axioms of probability. In the following we will describe what is a σ -algebra and what are the axioms of probability.

The choice of the probability space $(\Omega, \mathcal{F}, \mathbf{P}(\cdot))$ for a given random experiment is neither unique nor arbitrary. One of our main goals in the following is to demonstrate via examples how one can choose the probability space for given random experiment.

2.1 The Set of All Outcomes: The Sample Space Ω

The samples space is the set of all possible outcomes of the random experiment. Hence the sample space serves as the universal set for the outcomes of the random experiment. The particular outcomes, i.e., elements of Ω , are often denoted by the dummy variable ω .

The outcome of the random experiment will always be a single element of the sample space. Hence ω 's are mutual exclusive —i.e., when one outcome is observed no other outcome can be observed— and collectively exhaustive —i.e., at least one the outcomes will always be observed.

Example. For the random experiment of a single coin flip, the sample space can be chosen to be $\Omega = \{H, T\}$.

Example. For the random experiment of two coin flips, the sample space can be chosen to be $\Omega = \{HH, HT, TH, TT\}$.

Example. For the random experiment of n-coin flips, the sample space can be chosen to be $\Omega = \{(x_1, x_2, \dots, x_n) : x_j \in \{H, T\} \forall j \in \{1, 2, \dots, n\}\}$.

Example. For the random experiment composed of a single coin flip and a single observation of the whether for rain, the sample space can be chosen to be $\Omega = \{(H, \text{rain}), (H, \text{no-rain}), (T, \text{rain}), (T, \text{no-rain})\}$.

Example. For the random experiment of measuring the temperature in Kelvins the sample space can be chosen to be $\Omega = \{x \in \mathbb{R} : x \geq 0\}$.

Example. For the random experiment of measuring the temperature in Kelvins with a rounding to the closest integer, the sample space can be chosen to be $\Omega = \{x \in \mathbb{Z} : x \geq 0\}$.

2.2 The Set of All Events: The σ -algebra \mathcal{F}

Definition 2.1. A collection \mathcal{F} of subsets of Ω is a σ -algebra iff it satisfies the following three conditions.

- (i) $\Omega \in \mathcal{F}$.
- (ii) If $A \in \mathcal{F}$, then $\Omega \setminus A \in \mathcal{F}$.
- (iii) If $A_j \in \mathcal{F}$ for all $j \in \mathbb{Z}_+$, then $\bigcup_{j \in \mathbb{Z}_+} A_j \in \mathcal{F}$.

Remark. The condition (iii) tells us that by taking countable unions of elements of \mathcal{F} we cannot get an element outside \mathcal{F} . In other words \mathcal{F} is closed under countable unions. This requirement is stronger than being closed under finite unions and weaker than being closed under arbitrary unions. Thus a σ -algebra \mathcal{F} is always closed under finite unions but not it might not be closed under arbitrary unions.

HW. Show that being closed under countable unions implies being closed under finite unions.

HW. Show that if Ω is finite then being closed under finite unions implies being closed under countable unions.

HW. Show that for any set Ω , its power set 2^Ω is a σ -algebra of subset of Ω

Remark. We can apply De Morgan's Law to (iii) because of (i) and (ii). Thus \mathcal{F} is closed not only under countable unions, but also under countable intersections. It is not hard to see that \mathcal{F} is closed under countably many union, intersection, complementation operations applied in any order.

In probability space $(\Omega, \mathcal{F}, \mathbf{P}(\cdot))$, the σ -algebra \mathcal{F} is the set of all observable elements of 2^Ω (subsets of Ω) in the sense that the probability model is required to determine the probability of every single one of the elements of \mathcal{F} . Note that the probability model can be mute about the probabilities of the elements of the power set 2^Ω (the subsets of Ω) that are not elements of \mathcal{F} .

An event $A \in \mathcal{F}$ is said to occur if the outcome of the experiment ω is in A .

The sample space Ω is also called the certain event.

The empty set \emptyset is also called the impossible event.

Example 2.1. For the random experiment of two coin flips, let us assume the sample space to be $\Omega = \{HH, HT, TH, TT\}$.

- (a) If only the result of the first coin flip is of interest, then one can work with the following σ -algebra $\mathcal{F}_1 = \{\emptyset, \Omega, \{HH, HT\}, \{TH, TT\}\}$
- (b) If only the result of the second coin flip is of interest, then one can work with the following σ -algebra $\mathcal{F}_2 = \{\emptyset, \Omega, \{HH, TH\}, \{HT, TT\}\}$
- (c) If one is only interested in whether or not the result of the first and second coin flips are the same or not, then one can work with the σ -algebra $\mathcal{F}_e = \{\emptyset, \Omega, \{HH, TT\}, \{HT, TH\}\}$
- (d) If only the total number of H 's in the two coin flips are of interest, then one can work with the σ -algebra $\mathcal{F}_H = \{\emptyset, \Omega, \{HH\}, \{HT, TH\}, \{TT\}, \{HH, HT, TH\}, \{HT, TH, TT\}, \{HH, TT\}\}$.
- (e) If one is interested any two of results considered in (a), (b), and (c), then the only σ -algebra she can work with is 2^Ω .
- (f) If one is interested any two of results considered in (a), (b), and (d), then the only σ -algebra he can work with is 2^Ω .

Definition 2.2. A singleton of Ω , is a subset of Ω , composed of a single element, i.e. a set of the form $\{\omega\}$ for some $\omega \in \Omega$.

Remark. For most cases of interest, one can assume without loss of generality that all singleton are elements of the σ -algebra of the probability space, i.e., $\{\omega\} \in \mathcal{F}$ for all $\omega \in \Omega$.

HW. Show that if Ω is a countable set and $\{\omega\} \in \mathcal{F}$ for all $\omega \in \Omega$, then $\mathcal{F} = \mathcal{P}^\Omega$

Example 2.2. For the random experiment of measuring the temperature in Kelvins let us assume the sample space to be $\Omega = \{x \in \mathbb{R} : x \geq 0\}$. Furthermore, let us assume that we want open intervals of with rational end points to be events for our probability space, i.e. we assume that¹

$$\{(a, b) : a < b \text{ and } a, b \in \mathbb{Q}_+\} \subset \mathcal{F}.$$

The smallest σ -algebra satisfying this constraint is called the Borel σ -algebra. There are other equivalent characterizations but a detailed discussion of these characterizations is beyond the scope of an undergraduate probability course in an engineering department. Nevertheless the following observations justify why one might want to work with Borel σ -algebra.

(a) The open interval $(a, b) \in \mathcal{F}$ for any $a, b \in \mathbb{R}_{\geq 0}$ satisfying $a < b$.

Proof. For any such a and b there exists a decreasing sequence of rational numbers $\{a_j\}_{j \in \mathbb{Z}_+}$ such that $a_j \downarrow a$ and an increasing sequence of rational numbers $\{b_j\}_{j \in \mathbb{Z}_+}$ such that $b_j \uparrow b$, furthermore one can assume without loss of generality $a_1 < b_1$. Then $(a_i, b_i) \in \mathcal{F}$ by the hypothesis and $\bigcup_{i \in \mathbb{Z}_+} (a_i, b_i) = (a, b)$ by construction. Thus $(a, b) \in \mathcal{F}$ by (iii). \square

(b) **HW.** The open interval $(a, \infty) \in \mathcal{F}$ for any $a \in \mathbb{R}_{\geq 0}$.

(c) The singleton $\{a\} \in \mathcal{F}$ for any $a \in \mathbb{R}_{\geq 0}$.

Proof. For $\{0\}$ note that it is the complement of $(0, \infty)$ and thus $\{0\} \in \mathcal{F}$ by (ii) because $(0, \infty) \in \mathcal{F}$. For positive values of a , note that $\{a\} = \bigcap_{k=\lfloor 1/a \rfloor}^{\infty} (a - \frac{1}{k}, a + \frac{1}{k})$. Then $\{a\} \in \mathcal{F}$ because it is obtain by a countable intersections of elements of \mathcal{F} . \square

(d) **HW.** $[a, b), (a, b], [a, b] \in \mathcal{F}$ for any $a, b \in \mathbb{R}_{\geq 0}$ satisfying $a < b$.

One can show that the Borel σ -algebra includes all open and closed subsets of Ω . Borel σ -algebra, however, does not include all subsets of Ω , thus it is not equal to the power set \mathcal{P}^Ω .

Remark. For probability spaces with uncountable Ω there are technical issues that prevents us to from choosing the \mathcal{F} of the probability space to be \mathcal{P}^Ω . Nevertheless, in vast majority of such cases the \mathcal{F} we need for the probability space will a Borel σ -algebra. We will elaborate more on this issue after introducing the probability law and continuous probability models.

¹With a slight abuse of notation we denote by (a, b) not the ordered pair but the set of real numbers in the open interval with the end points a and b .

2.3 Probability Law and Probability Axioms

In a given probability space, the sample space specifies the possible outcomes of random experiment. The σ -algebra specifies the events that probability model will assign probabilities to. The probability law specifies the probabilities of the events. In order to assign probabilities to the events in a consistent way probability law need to satisfy the probability axioms first put forward by Andrey Kolmogorov in 1933. Thus the probability axioms are sometimes called Kolmogorov's axioms.

2.3.1 Probability Axioms

(A1) **Non-negativity:** $\mathbf{P}(A) \geq 0$ for all events A .

(A2) **Normalization:** $\mathbf{P}(\Omega) = 1$.

(A3) **σ -additivity (Countable Additivity):** If A_1, A_2, \dots is a disjoint sequence of events, i.e. $A_i \cap A_j = \emptyset$ for all positive integers i and j such that $i \neq j$, then

$$\mathbf{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbf{P}(A_i) \quad (2.1)$$

The probability axioms essentially tells us that the probability law should be a σ -additive function of the form $\mathbf{P}(\cdot) : \mathcal{F} \rightarrow [0, 1]$ satisfying $\mathbf{P}(\Omega) = 1$. The probability axioms, however, do not state all of the identities satisfied by all probability laws, e.g. " $\mathbf{P}(\emptyset) = 0$ " is not a probability axiom but it is satisfied by all probability laws. The following is not a exhaustive list of properties of probability laws; it is a partial list of some of the more important ones. Nevertheless, following the proofs of these statements will guide you how to prove such relations.

2.3.2 Properties of Probability Laws

(P1) $\mathbf{P}(\emptyset) = 0$

Proof. Let us consider the events $A_1 = \Omega$ and $A_{1+j} = \emptyset$ for all $j \in \mathbb{Z}_+$. Then $A_j \cap A_i = \emptyset$ for all $i \neq j$ and $\Omega = \bigcup_{i=1}^{\infty} \emptyset$.

$$\begin{aligned} \mathbf{P}(\Omega) &= \mathbf{P}(\Omega) + \sum_{i=2}^{\infty} \mathbf{P}(\emptyset) && \text{by (A3)} \\ 0 &= \lim_{j \rightarrow \infty} (j-1)\mathbf{P}(\emptyset) && \text{because } \sum_{i=2}^{\infty} \mathbf{P}(\emptyset) = \lim_{j \rightarrow \infty} \sum_{i=2}^j \mathbf{P}(\emptyset). \end{aligned}$$

Thus $\mathbf{P}(\emptyset) = 0$. □

(P2) If A_1, A_2, \dots, A_n are disjoint events then $\mathbf{P}(\bigcup_{i=1}^n A_i) = \sum_{i=1}^n \mathbf{P}(A_i)$.

Proof. Let $A_i = \emptyset$ for all $i > n$, then A_i 's are disjoint events and $\bigcup_{i=1}^n A_i = \bigcup_{i=1}^{\infty} A_i$. Thus

$$\begin{aligned} \mathbf{P}(\bigcup_{i=1}^n A_i) &= \mathbf{P}(\bigcup_{i=1}^{\infty} A_i) && \text{by } \bigcup_{i=1}^n A_i = \bigcup_{i=1}^{\infty} A_i \\ &= \sum_{i=1}^{\infty} \mathbf{P}(A_i) && \text{by (A3)} \\ &= \sum_{i=1}^n \mathbf{P}(A_i) && \text{because } \mathbf{P}(A_j) = \mathbf{P}(\emptyset) \text{ by construction and } \mathbf{P}(\emptyset) = 0. \end{aligned}$$

□

Remark. The finite additivity property (P2) is proved by using the countable additivity axiom (A3). Thus the countable additivity axiom (A3) implies the finite additivity property (P2). However, the converse statement is not true: the finite additivity property (P2) does not imply the countable additivity axiom (A3).

(P3) $\mathbf{P}(A^c) = 1 - \mathbf{P}(A)$ for any event A .

Proof. For any event A we have $A \cap A^c = \emptyset$.

$$\begin{aligned} \mathbf{P}(A) + \mathbf{P}(A^c) &= \mathbf{P}(A \cup A^c) && \text{by finite additivity, i.e. by (P2)} \\ &= 1 && \text{by (A2) because } A \cup A^c = \Omega \end{aligned}$$

□

(P4) $\mathbf{P}(A - B) = \mathbf{P}(A) - \mathbf{P}(A \cap B)$ for all events A and B .

Proof. $A = (A \cap B) \cup (A \cap B^c)$ and $(A \cap B) \cap (A \cap B^c) = \emptyset$ for any pair of events A and B . Thus as a result of finite additivity, i.e., (P2) we have

$$\mathbf{P}(A) = \mathbf{P}(A \cap B) + \mathbf{P}(A \cap B^c).$$

Thus the identity follows from $A - B = A \cap B^c$. □

(P5) **HW.** $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B)$ for all events A and B .

(P6) **HW.** $\mathbf{P}(A \cup B) \leq \mathbf{P}(A) + \mathbf{P}(B)$ for any events A and B .

(P7) **HW.** If $B \subset A$ for two event B and A , then $\mathbf{P}(B) \leq \mathbf{P}(A)$.

(P8) For any countable collection of events A_1, A_2, \dots

$$\mathbf{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \mathbf{P}(A_1) + \sum_{i=2}^{\infty} \mathbf{P}\left(A_i \cap \left(\bigcap_{j=1}^{i-1} A_j^c\right)\right) \quad (2.2)$$

$$\leq \sum_{i=1}^{\infty} \mathbf{P}(A_i) \quad (2.3)$$

Proof. Note that $\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} B_i$ where $B_1 = A_1$ and $B_i = A_i \cap \left(\bigcup_{j=1}^{i-1} A_j\right)^c$ for all $i \in \{2, 3, \dots\}$. As a result of De Morgan law we have $B_i = A_i \cap \left(\bigcap_{j=1}^{i-1} A_j^c\right)$ and thus B_i 's are disjoint sets and (2.2) follows from the finite additivity property (P2). (2.3) follows from (2.2) via (P7) because $A_i \cap \left(\bigcap_{j=1}^{i-1} A_j^c\right) \subset A_i$. □

(P9) **HW.**

$$\mathbf{P}(A_1 \cup A_2 \cup A_3) = \mathbf{P}(A_1) + \mathbf{P}(A_2) + \mathbf{P}(A_3) - \mathbf{P}(A_1 \cap A_2) - \mathbf{P}(A_2 \cap A_3) - \mathbf{P}(A_3 \cap A_1) + \mathbf{P}(A_1 \cap A_2 \cap A_3)$$

2.4 Discrete Probability Models

Definition 2.3. A probability space $(\Omega, \mathcal{F}, \mathbf{P}(\cdot))$ is a *discrete probability space* iff Ω is countable.

In discrete probability spaces the σ -algebra \mathcal{F} is almost always 2^Ω . When the sample space Ω not only countable but also finite, the following mapping of 2^Ω to the interval $[0, 1]$ satisfies axioms of probability, resulting probability law is called *discrete uniform law*.

$$\mathbf{P}(A) = \frac{|A|}{|\Omega|} \quad \forall A \in 2^\Omega.$$

In some discrete probability spaces, Ω can be expressed as a Cartesian product of the form $\Omega = \Omega_1 \times \dots \times \Omega_n$, i.e.

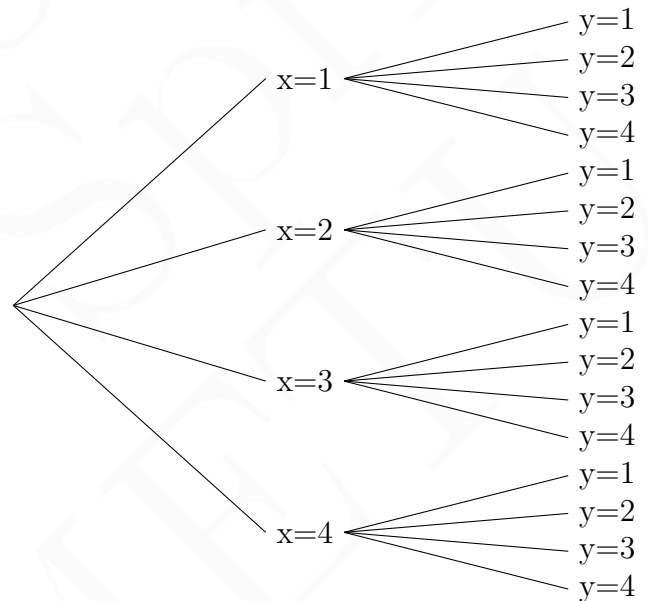
$$\Omega = \{(\omega_1, \dots, \omega_n) : \omega_j \in \Omega_j \quad \forall j \in \{1, \dots, n\}\}.$$

If n is a small number one can use trees and if $n = 2$ one can use tables to describe these sample spaces.

Example 2.3. Let us consider two rolls of a tetrahedral die and assume that $\mathcal{F} = 2^\Omega$ and $\mathbf{P}(\cdot)$ is the discrete uniform law. Then the sample space Ω is given by

$$\Omega = \{(x, y) : x \in \{1, 2, 3, 4\} \text{ and } y \in \{1, 2, 3, 4\}\} \tag{2.4}$$

	x=1	x=2	x=3	x=4
y=1				
y=2				
y=3				
y=4				



Let us denote the outcome of the first roll of the die by X and the second roll of the die by Y .

$$\begin{aligned} \mathbf{P}(\{(x, y) \in \Omega : x = 1\}) &= \frac{4}{16} & \mathbf{P}(\{(x, y) \in \Omega : x \wedge y = 2\}) &= \frac{5}{16} \\ \mathbf{P}(\{(x, y) \in \Omega : x = y\}) &= \frac{4}{16} & \mathbf{P}(\{(x, y) \in \Omega : x + y \text{ is odd}\}) &= \frac{8}{16} \\ \mathbf{P}(\{(x, y) \in \Omega : x > y\}) &= \frac{6}{16} & \mathbf{P}(\{(x, y) \in \Omega : x = 4 \text{ or } y = 4\}) &= \frac{7}{16} \end{aligned}$$

HW. For two rolls of a tetrahedral die if we want “the sum of the integers we get in two rolls being even” and “the integers we get in two rolls being the same” to be events, what is the size of smallest set of all possible events (i.e. smallest σ -algebra)?

HW. For two rolls of a tetrahedral die if we want “the sum of the integers we get in two rolls being even” and “the sum of the two rolls being larger than or equal to 6” to be events, what is the size of smallest set of all possible events (i.e. smallest σ -algebra)?

Example 2.4. Forty candidates are to take a driver's license exam. The probability that at least one of the candidates fails is 0.2. The probability that at least two of the candidates fail is 0.15. Let the number of candidates who fail be the outcome of the experiment.

- What is the sample space of the experiment?
- What is the probability of the event that all forty candidates will pass?
- What is the probability of the event that exactly one of the candidates will fail?
- What is the probability of the event that at most one of the candidates will fail?
- What is the smallest set of all events (the smallest σ -algebra) in which we can express all the information we are given?

a) $\Omega = \{0, 1, \dots, 40\}$

Remark. $\{0, 1, \dots, 40\}$ is the smallest sample space we can use. We can choose any superset of $\{0, 1, \dots, 40\}$ to be the sample space but outcomes outside $\{0, 1, \dots, 40\}$ will never happen.

- b) All forty candidates pass whenever the event $\{\omega \in \Omega : \omega = 0\}$, i.e., $\{0\}$, occurs.

$$\begin{aligned} \mathbf{P}(\{0\}) &= 1 - \mathbf{P}(\{\omega \in \Omega : \omega \geq 1\}) \\ &= 1 - 0.2 \\ &= 0.8 \end{aligned}$$

- c) Exactly one candidate fails whenever the event $\{\omega \in \Omega : \omega = 1\}$, i.e., $\{1\}$, occurs.

$$\begin{aligned} \mathbf{P}(\{1\}) &= \mathbf{P}(\{\omega \in \Omega : \omega \geq 1\} \setminus \{\omega \in \Omega : \omega \geq 2\}) \\ &= \mathbf{P}(\{\omega \in \Omega : \omega \geq 1\}) - \mathbf{P}(\{\omega \in \Omega : \omega \geq 2\}) \\ &= 0.2 - 0.15 \\ &= 0.05 \end{aligned}$$

- d) At most one candidate fails whenever the event $\{\omega \in \Omega : \omega \leq 1\}$, i.e., $\{0, 1\}$, occurs.

$$\begin{aligned} \mathbf{P}(\{0, 1\}) &= \mathbf{P}(\{0\}) + \mathbf{P}(\{1\}) \\ &= 0.8 + 0.05 \\ &= 0.85 \\ \mathbf{P}(\{0, 1\}) &= 1 - \mathbf{P}(\{\omega \in \Omega : \omega \geq 2\}) \\ &= 1 - 0.15 \\ &= 0.85 \end{aligned}$$

- e) We need to have a σ -algebra in which both $\{\omega \in \Omega : \omega \geq 1\}$ and $\{\omega \in \Omega : \omega \geq 2\}$ are events²

$$\mathcal{F} = \{\emptyset, \Omega, \{\omega \in \Omega : \omega \geq 1\}, \{\omega \in \Omega : \omega \geq 2\}, \{0\}, \{0, 1\}, \{1\}, \{\omega \in \Omega : \omega \neq 1\}\}.$$

Remark. Note that one can make calculations made in the previous parts of the problem by assuming that the σ -algebra, i.e. the set of all events, is the power set 2^Ω . Furthermore, this is what is done in most problems without explicitly mentioning it when working with discrete probability models.

²Note that $\{\{0\}, \{1\}, \{\omega \in \Omega : \omega \geq 2\}\}$ forms a partition of the sample space Ω , and all elements of \mathcal{F} except for \emptyset , can be represented as the unions of the elements of this partition. For countable sample spaces this is always the case, i.e. for any $\mathcal{F} \subset 2^\Omega$ there exists a unique partition $\{B_j\}_{j \in J}$ of Ω such that $\mathcal{F} = \{\cup_{i \in I} B_i : I \subset J\}$ where $\cup_{i \in \emptyset} B_i$ stands for \emptyset .

Example 2.5. A candidate takes the driver's license exam until he passes. The probability that he passes the exam in his i^{th} attempt is 3^{-i} .

- What is the sample space of the experiment?
- What is the probability that candidate will get a drivers license?
- What is the probability that candidate will pass the exam at an odd numbered attempt?

(a) Either the candidate will pass the exam in one his attempts or he will fail in all of them.

$$\Omega = \{F, 1, 2, \dots\}$$

(b) Candidate will get a drivers license iff the event $\{\omega \in \mathbb{Z}_+\}$ occurs.

$$\begin{aligned} \mathbf{P}(\{\omega \in \mathbb{Z}_+\}) &= \sum_{\omega \in \mathbb{Z}_+} \mathbf{P}(\{\omega\}) && \text{by } \sigma\text{-additivity} \\ &= \sum_{i \in \mathbb{Z}_+} 3^{-i} \\ &= \frac{1}{3} \sum_{j=0}^{\infty} 3^{-j} \\ &= \frac{1}{3} \frac{1}{1-1/3} \\ &= 0.5 \end{aligned}$$

(c) Candidate will pass the exam at an odd numbered attempt iff the event $\{\frac{\omega+1}{2} \in \mathbb{Z}_+\}$ occurs.

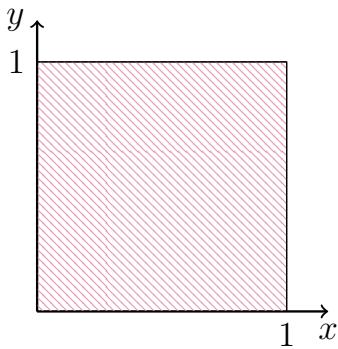
$$\begin{aligned} \mathbf{P}(\{\omega \in \Omega : \frac{\omega+1}{2} \in \mathbb{Z}_+\}) &= \sum_{\omega \in \Omega: \frac{\omega+1}{2} \in \mathbb{Z}_+} \mathbf{P}(\{\omega\}) && \text{by } \sigma\text{-additivity} \\ &= \sum_{i \in \mathbb{Z}_+} \mathbf{P}(\{(2i-1)\}) \\ &= \sum_{i \in \mathbb{Z}_+} 3^{-(2i-1)} \\ &= \frac{1}{3} \sum_{j=0}^{\infty} 3^{-2j} \\ &= \frac{1}{3} \frac{1}{1-1/9} \\ &= 3/8 \end{aligned}$$

2.5 Continuous Probability Models

The sample space of some probability spaces is uncountable, e.g., an open set in n -dimensional Euclidean space \mathbb{R}^n . The set of all events (i.e., the σ -algebra) in such a probability space is usually the corresponding Borel σ -algebra —i.e., the smallest σ -algebra of subsets of Ω in which all open sets are elements/events. The probability law in a such model is usually described via a density function that tells the density of the probability per unit length/area/volume at different parts of the sample space. Such a model is called a continuous probability model.³ In continuous probability models, we will only calculate the probabilities of subsets of Ω that you have already learned how to take integrals over using the Riemann integral in your freshman-year calculus course. This collection of subsets Ω form a subset of the Borel σ -algebra that is sufficient to analyze and understand vast majority of the random experiments that are modeled using continuous probability models.

³In the above description, we implicitly assume that we know which subsets of Ω are open. When Ω is an open set in \mathbb{R}^n , the open subsets of Ω are those that are open sets in \mathbb{R}^n . More generally, the uncountable sample space Ω will be a subset of a space \mathcal{T} with a given collection of open sets, and the open subsets of Ω will be the intersections of those sets with Ω .

Example 2.6.

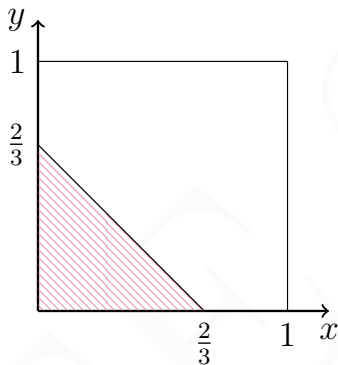


$$\Omega = \{(x, y) : x \in (0, 1) \text{ and } y \in (0, 1)\}$$

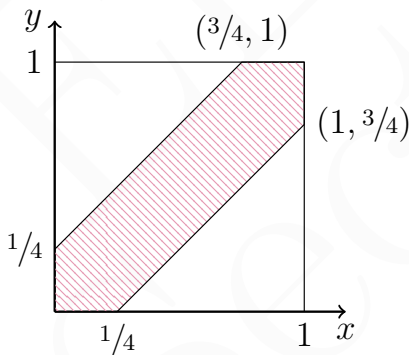
\mathcal{F} is all subsets of Ω that can be obtained from open set in Ω via countable intersection, union, and complementation operations.

$$\mathbf{P}(A) = \text{the area of the event } A \quad \forall A \in \mathcal{F}$$

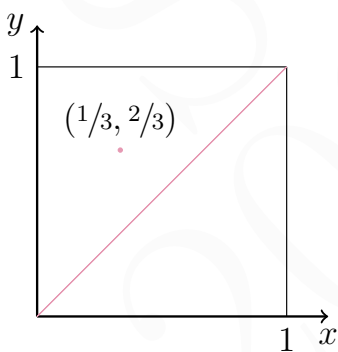
Note that the probability law is uniform in the sense that it is just proportional to the area; it does not depend on the place or shape of A in Ω .



$$\begin{aligned} \mathbf{P}\left(\left\{(x, y) \in \Omega : x + y < \frac{2}{3}\right\}\right) &= \frac{1}{2} \cdot \frac{2}{3} \cdot \frac{2}{3} \\ &= \frac{2}{9} \end{aligned}$$



$$\begin{aligned} \mathbf{P}\left(\left\{(x, y) \in \Omega : |x - y| < \frac{1}{4}\right\}\right) &= 1 - 2 \left(\frac{1}{2} \cdot \frac{3}{4} \cdot \frac{3}{4}\right) \\ &= \frac{7}{16} \end{aligned}$$

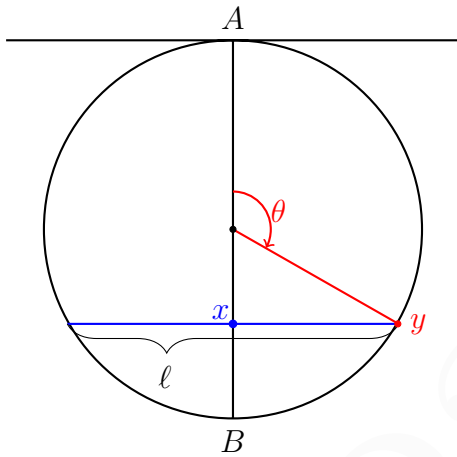


$$\begin{aligned} \mathbf{P}(\{(1/3, 2/3)\}) &= 0 \\ \mathbf{P}(\{(x, y) \in \Omega : x \in \mathbb{Q} \text{ and } y \in \mathbb{Q}\}) &= 0 \quad \text{by } \sigma\text{-additivity} \\ \mathbf{P}(\{(x, y) \in \Omega : |x - y| = 0\}) &= 0 \\ \mathbf{P}(\{(x, y) \in \Omega : |x - y| \in \mathbb{Q}\}) &= 0 \quad \text{by } \sigma\text{-additivity} \end{aligned}$$

Remark 2.1. The area of any line segment is zero. Thus one can replace the sample space Ω with the sample space $\tilde{\Omega}$ given below without changing the probability space in any essential way.

$$\tilde{\Omega} = \{(x, y) : x \in [0, 1] \text{ and } y \in [0, 1]\}$$

Example 2.7.



Let us consider the random experiment of choosing chords of a circle that are parallel to a given line “uniformly at random.” We are interested in finding the probability of the event that the length of the chord being at least $\sqrt{3}$ times the radius of the circle:

$$\mathbf{P}\left(\left\{\ell \geq \sqrt{3}r\right\}\right) = ?$$

- If we are choosing the midpoints of the chord x uniformly on the line segment AB then only midpoint that are at most $r/2$ away from the center of the circle will satisfy the condition $\ell \geq \sqrt{3}r$. Then

$$\mathbf{P}\left(\left\{\ell \geq \sqrt{3}r\right\}\right) = \frac{1}{2r}r = \frac{1}{2}$$

- If the endpoint y or equivalently the angle θ is chosen uniformly over the possible values then only those values of θ satisfying the condition $\sin \theta \geq \frac{\sqrt{3}}{2}$ will satisfy the condition $\ell \geq \sqrt{3}r$. Then

$$\mathbf{P}\left(\left\{\ell \geq \sqrt{3}r\right\}\right) = \mathbf{P}\left(\left\{\sin \theta \geq \frac{\sqrt{3}}{2}\right\}\right) = \mathbf{P}\left(\left\{\theta \in \left[\frac{\pi}{3}, \frac{2\pi}{3}\right]\right\}\right) = \frac{1}{3}$$

Remark 2.2. Note that both for discrete and continuous models the procedure we have employed can be summarized as follows.

- Step 1** Specify a sample space Ω for the experiment.
- Step 2** Specify⁴ a probability law $\mathbf{P}(\cdot)$.
- Step 3** Identify an event of interest.
- Step 4** Calculate.

Note that calculation in step four is necessary because $\mathbf{P}(\cdot)$ is usually specified either partially or implicitly.

⁴This step implicitly specifies the set of all events \mathcal{F} ; we do not deal with it explicitly because we are interested in determining the probability of only one or a few elements of \mathcal{F} .

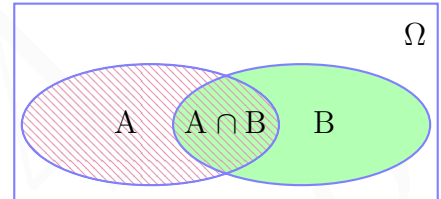
3 Conditional Probability

The probability law can be interpreted as an expression of our belief about the outcome of a random experiment. If we are provided some partial information about the outcome of the experiment then our beliefs about the outcome will change. We reflect this formally using conditional probabilities.

3.1 Conditional Probability of An Event Given Another Event

Definition 3.1. Let $(\Omega, \mathcal{F}, \mathbf{P}(\cdot))$ be a probability space and B be an event with positive probability, i.e., $\mathbf{P}(B) > 0$. Then the conditional probability any event A given event B is

$$\mathbf{P}(A|B) := \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)}. \tag{3.1}$$



The conditional probability $\mathbf{P}(A|B)$ is undefined for the case when $\mathbf{P}(B) = 0$.

Example 3.1. Recall the two rolls of fair a tetrahedral die considered in Example 2.3.

$$\begin{aligned} B &= \{(x, y) \in \Omega : x \wedge y = 2\} \\ A_m &= \{(x, y) \in \Omega : x \vee y = m\} && m \in 1, 2, 3, 4 \\ \mathbf{P}(A_m|B) &=? && m \in 1, 2, 3, 4 \end{aligned}$$

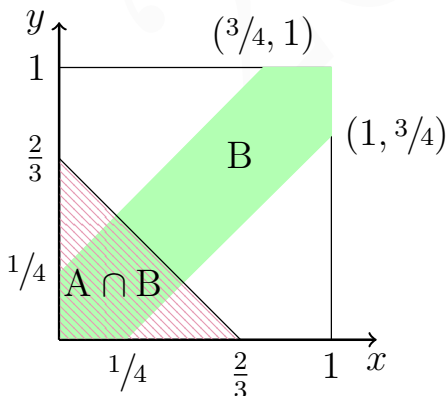
Since $\mathbf{P}(\cdot)$ is the discrete uniform law we can express the conditional probabilities as

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

	x=1	x=2	x=3	x=4
y=1	A ₁	A ₂	A ₃	A ₄
y=2	A ₂	A ₂	A ₃	A ₄
y=3	A ₃	A ₃	A ₃	A ₄
y=4	A ₄	A ₄	A ₄	A ₄

Example 3.2. Recall the continuous uniform probability law considered in Example 2.6

$$\begin{aligned} \Omega &= \{(x, y) : x \in (0, 1) \text{ and } y \in (0, 1)\} \\ \mathbf{P}(E) &= \text{the area of the event } E && \forall E \in \mathcal{F} \\ A &= \{(x, y) \in \Omega : x + y < 2/3\} \\ B &= \{(x, y) \in \Omega : |x - y| < 1/4\} \\ \mathbf{P}(A|B) &=? \end{aligned}$$



$$\begin{aligned} \mathbf{P}(B) &= \frac{7}{16} \\ \mathbf{P}(A \cap B) &= \mathbf{P}(A) - 2 \left(\frac{1}{2} \left(\frac{2}{3} - \frac{1}{4} \right) \left[\frac{1}{2} \left(\frac{2}{3} - \frac{1}{4} \right) \right] \right) \\ &= \frac{2}{9} - \frac{1}{2} \left(\frac{5}{12} \right)^2 = \frac{13}{32 \cdot 3} \\ \mathbf{P}(A|B) &= \frac{13}{42} \end{aligned}$$

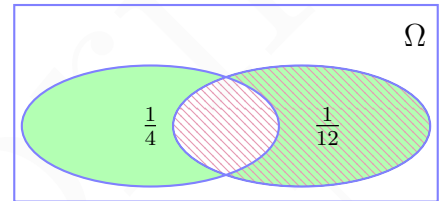
Example 3.3. A candidate takes the driver’s license exam until he passes. The probability that he passes the exam in his i^{th} attempt is 3^{-i} . Given that he passes the exam at an odd numbered attempt what is the conditional probability that he passes the exam in one of his first 4 attempts.

For the sample space $\Omega = \{F, 1, 2, \dots\}$ introduced in Example 2.5

$$\begin{aligned}
 A &= \{1, 2, 3, 4\} \\
 B &= \{\omega \in \Omega : \frac{\omega+1}{2} \in \mathbb{Z}_+\} \\
 \mathbf{P}(A \cap B) &= \mathbf{P}(\{1, 3\}) &&= \frac{10}{27} \\
 \mathbf{P}(B) &= \frac{3}{8} &&\text{see Example 2.5} \\
 \mathbf{P}(A|B) &= \frac{80}{81}
 \end{aligned}$$

Example 3.4. A family is contemplating getting a dog, cat, or both. They get a dog with $2/3$ probability, a cat with $1/2$ probability, and a dog or a cat with $3/4$ probability. Given that they decided to get a single pet, what is the probability that their pet is a cat?

$$\begin{aligned}
 \mathbf{P}(D) &= \frac{2}{3} && \mathbf{P}(C) = \frac{1}{2} \\
 \mathbf{P}(D \cup C) &= \frac{3}{4} && B = (D \setminus C) \cup (C \setminus D) \\
 \mathbf{P}(D \setminus C) &= \mathbf{P}(D \cup C) - \mathbf{P}(C) &&= \frac{1}{4} \\
 \mathbf{P}(C \setminus D) &= \mathbf{P}(D \cup C) - \mathbf{P}(D) &&= \frac{1}{12} \\
 \mathbf{P}(C|B) &= \frac{1/12}{1/4 + 1/12} &&= \frac{1}{4}
 \end{aligned}$$



3.2 Conditional Probability Law

Theorem 3.1. Let $(\Omega, \mathcal{F}, \mathbf{P}(\cdot))$ be a probability space and B be an event with positive probability, i.e., $\mathbf{P}(B) > 0$. Then $\mathbf{P}(\cdot|B)$ is a probability law (i.e., $\mathbf{P}(\cdot|B)$ satisfies axioms of probability) and hence $(\Omega, \mathcal{F}, \mathbf{P}(\cdot|B))$ is a probability space.

Proof. (i) $\mathbf{P}(A|B) \geq 0$ for all $A \in \mathcal{F}$ by (3.1) because $\mathbf{P}(A \cap B) \geq 0$ and $\mathbf{P}(B) > 0$.

(ii)

$$\mathbf{P}(\Omega|B) = \frac{\mathbf{P}(\Omega \cap B)}{\mathbf{P}(B)} = \frac{\mathbf{P}(B)}{\mathbf{P}(B)} = 1$$

(iii) For any countable collection of disjoint sets A_1, A_2, \dots

$$\begin{aligned}
 \mathbf{P}\left(\bigcup_{j=1}^{\infty} A_j \middle| B\right) &= \frac{\mathbf{P}\left(\left(\bigcup_{j=1}^{\infty} A_j\right) \cap B\right)}{\mathbf{P}(B)} && (3.1), \\
 &= \frac{\mathbf{P}\left(\bigcup_{j=1}^{\infty} (A_j \cap B)\right)}{\mathbf{P}(B)} && \text{by the distributivity of } \cap \text{ on } \cup, \\
 &= \sum_{j=1}^{\infty} \frac{\mathbf{P}(A_j \cap B)}{\mathbf{P}(B)} && \text{by the } \sigma\text{-additivity of } \mathbf{P}(\cdot), \\
 &= \sum_{j=1}^{\infty} \mathbf{P}(A_j|B) && (3.1).
 \end{aligned}$$

□

The conditional probability law $\mathbf{P}(\cdot|B)$ can also be interpreted as the probability law of the probability space $(B, \mathcal{F}_B, \mathbf{P}(\cdot|B))$ for the set of all event⁵ $\mathcal{F}_B := \{A \in \mathcal{F} : A \subset B\}$.

3.3 Multiplication Rule (Chain Rule) and Modeling

If $\mathbf{P}\left(\bigcap_{j=1}^n A_j\right) > 0$, then

$$\mathbf{P}\left(\bigcap_{j=1}^n A_j\right) = \mathbf{P}(A_1) \prod_{j=2}^n \mathbf{P}\left(A_j \mid \bigcap_{i=1}^{j-1} A_i\right) \tag{3.2}$$

Why?

$$\begin{aligned} \mathbf{P}\left(\bigcap_{j=1}^n A_j\right) &= \mathbf{P}\left(A_n \mid \bigcap_{j=1}^{n-1} A_j\right) \mathbf{P}\left(\bigcap_{j=1}^{n-1} A_j\right) && \text{by (3.1)} \\ &= \mathbf{P}\left(A_n \mid \bigcap_{j=1}^{n-1} A_j\right) \mathbf{P}\left(A_{n-1} \mid \bigcap_{j=1}^{n-2} A_j\right) \mathbf{P}\left(\bigcap_{j=1}^{n-2} A_j\right) && \text{by (3.1)} \\ &\vdots \\ &= \left(\prod_{j=2}^n \mathbf{P}\left(A_j \mid \bigcap_{i=1}^{j-1} A_i\right)\right) \mathbf{P}(A_1) \end{aligned}$$

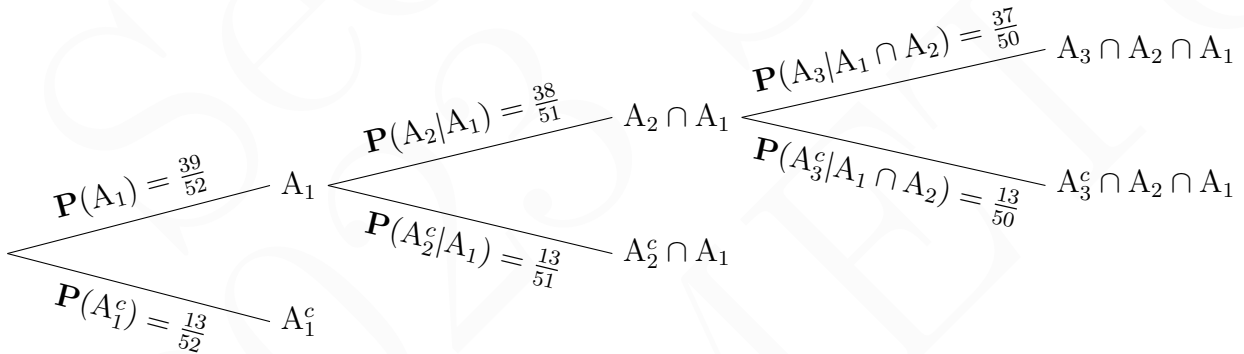
Example 3.5. Three cards are drawn from an ordinary 52 card deck without replacement, i.e.,

- drawn cards are not placed back,
- each card that has not been drawn yet is equally likely to be drawn at each step.

What is the probability that there are no hearts among the three cards drawn?

$$A_i = \{\text{the } i^{\text{th}} \text{ card is not a heart}\} \quad i = \mathbb{Z}_+$$

$$\mathbf{P}\left(\bigcap_{i=1}^3 A_i\right) = ?$$



$$\mathbf{P}(A_1) = \frac{39}{52}$$

$$\mathbf{P}(A_2|A_1) = \frac{38}{51} \quad \mathbf{P}(A_3|A_1 \cap A_2) = \frac{37}{50}$$

$$\begin{aligned} \mathbf{P}\left(\bigcap_{i=1}^3 A_i\right) &= \mathbf{P}(A_1) \mathbf{P}(A_2|A_1) \mathbf{P}(A_3|A_1 \cap A_2) \\ &= \frac{39}{52} \cdot \frac{38}{51} \cdot \frac{37}{50} \\ &= \frac{19 \cdot 37}{1700} \end{aligned}$$

⁵Note that the set of all events (the σ -algebra) \mathcal{F}_B can also be expressed as $\mathcal{F}_B = \{A \cap B : A \in \mathcal{F}\}$.

Example 3.6 (Radar Detection).

- (i) When an aircraft is present radar detects it with probability 0.99.
- (ii) When there are no aircraft, the radar generates a (false) alarm, with probability 0.10.
- (iii) An aircraft is present with probability 0.05.

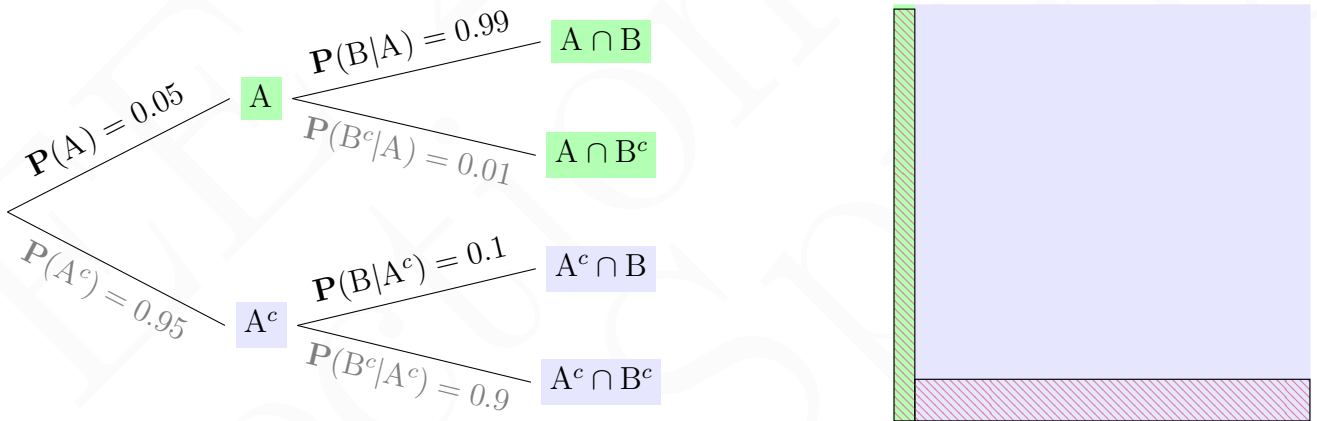
- a) What is the probability of a false alarm, i.e. $\mathbf{P}(\{\nexists \text{ aircraft and } \exists \text{ alarm}\}) = ?$
- b) What is the probability of a missed detection, i.e. $\mathbf{P}(\{\exists \text{ aircraft and } \nexists \text{ alarm}\}) = ?$

Let A be the event of having an aircraft and B be the event of radar generating an alarm, i.e.,

$$A = \{\exists \text{ aircraft}\} \quad \text{and} \quad B = \{\exists \text{ alarm}\}.$$

Then

$$\mathbf{P}(B|A) = 0.99, \quad \mathbf{P}(B|A^c) = 0.10, \quad \mathbf{P}(A) = 0.05.$$



In the figure on the right the alarm event, i.e. event B , is marked with the pattern. Probabilities of all events are proportional of the corresponding region.

- a) **False Alarm** $\{\nexists \text{ aircraft and } \exists \text{ alarm}\} = A^c \cap B$

$$\mathbf{P}(A^c \cap B) = \mathbf{P}(A^c) \mathbf{P}(B|A^c) = (1 - \mathbf{P}(A)) \mathbf{P}(B|A^c) = 0.95 \cdot 0.1 = 0.095$$

- b) **Missed Detection** $\{\exists \text{ aircraft and } \nexists \text{ alarm}\} = A \cap B^c$

$$\mathbf{P}(A \cap B^c) = \mathbf{P}(A) \mathbf{P}(B^c|A) = \mathbf{P}(A) (1 - \mathbf{P}(B|A)) = 0.05 \cdot 0.01 = 0.0005$$

How about $\mathbf{P}(B)$, $\mathbf{P}(A|B)$, and $\mathbf{P}(A|B^c)$?

$$\begin{aligned} \mathbf{P}(B) &= \mathbf{P}(A^c \cap B) + \mathbf{P}(A \cap B) \\ &= \mathbf{P}(A^c \cap B) + 0.05 \cdot 0.99 \\ &= 0.1445 \end{aligned}$$

$$\begin{aligned} \mathbf{P}(B^c) &= 1 - \mathbf{P}(B) \\ &= 0.8555 \end{aligned}$$

“Total probability theorem”

$$\begin{aligned} \mathbf{P}(A^c|B) &= \frac{\mathbf{P}(A^c \cap B)}{\mathbf{P}(B)} \\ &= \frac{0.095}{0.1445} \\ &\approx 0.657 \end{aligned}$$

$$\begin{aligned} \mathbf{P}(A|B) &= 1 - \mathbf{P}(A^c|B) \\ &= \frac{0.0495}{0.1445} \\ &\approx 0.342 \end{aligned}$$

“Bayes’ Rule”

$$\begin{aligned} \mathbf{P}(A|B^c) &= \frac{\mathbf{P}(A \cap B^c)}{\mathbf{P}(B^c)} \\ &= \frac{0.0005}{0.8555} \\ &\approx .0006 \end{aligned}$$

$$\begin{aligned} \mathbf{P}(A^c|B^c) &= 1 - \mathbf{P}(A|B^c) \\ &= \frac{0.855}{0.8555} \\ &\approx 0.9994 \end{aligned}$$

“Bayes’ Rule”

3.4 Total Probability Theorem

Let A_1, A_2, \dots, A_n be disjoint events satisfying⁶ both $\mathbf{P}(A_j) > 0$ for all $j \in \{1, 2, \dots, n\}$ and $\bigcup_{i=1}^n A_j = \Omega$, then for any event B , we have

$$\mathbf{P}(B) = \sum_{j=1}^n \mathbf{P}(A_j \cap B) \tag{3.3}$$

$$= \sum_{j=1}^n \mathbf{P}(A_j) \mathbf{P}(B|A_j) \tag{3.4}$$

Why?

$$\begin{aligned} \mathbf{P}(B) &= \mathbf{P}\left(\left(\bigcup_{j=1}^n A_j\right) \cap B\right) && \text{because } \bigcup_{j=1}^n A_j = \Omega, \\ &= \mathbf{P}\left(\bigcup_{j=1}^n (A_j \cap B)\right) && \text{by distributivity of intersection over unions,} \\ &= \sum_{j=1}^n \mathbf{P}\left((A_j \cap B)\right) && \text{by the additivity of the probability law,} \\ &= \sum_{j=1}^n \mathbf{P}(A_j) \mathbf{P}(B|A_j) && \text{by the the definition of conditional probability.} \end{aligned}$$

Remark 3.1. If we are interested in asserting total probability theorem for a particular even B only, then we can replace the hypothesis $\bigcup_{j=1}^n A_j = \Omega$ with $B \subset \bigcup_{j=1}^n A_j$.

Remark 3.2. Since the probability law is not only additive but also countably additive we can assert the same identity for any countable collection of disjoint sets, and hence for the case when “ $n = \infty$ ” as well.

Example 3.7. A fair tetrahedral die rolled once. If the result is either 1 or 2, then it is rolled once more, otherwise there is no second roll. What is probability that the sum of all the rolls is greater than or equal to 4.

$$\begin{aligned} A_i &= \{\text{The result of the first roll is } i\} && \forall i \in \{1, 2, 3, 4\} \\ B &= \{\text{The sum is at least 4}\} \\ \mathbf{P}(A_i) &= \frac{1}{4} && \forall i \in \{1, 2, 3, 4\} \\ \mathbf{P}(B|A_1) &= \frac{2}{4} && \mathbf{P}(B|A_2) = \frac{3}{4} \\ \mathbf{P}(B|A_3) &= 0 && \mathbf{P}(B|A_4) = 1 \\ \mathbf{P}(B) &= \frac{1}{4} \cdot \frac{2}{4} + \frac{1}{4} \cdot \frac{3}{4} + \frac{1}{4} \cdot 0 + \frac{1}{4} \cdot 1 && = \frac{9}{16} \end{aligned}$$

Example 3.8. [BT08] Alice is taking a probability class and at the end of each week she can be either up-to-date or she may have fallen behind. If she is up-to-date in a given week, the probability that she will be up-to-date (or behind) in the next week is 0.8 (or 0.2, respectively). If she is behind in a given week, the probability that she will be up-to-date (or behind) in the next week is 0.4 (or 0.6, respectively). Alice is (by default) up-to-date when she starts the class. What is the probability that she is up-to-date after i weeks? What is the probability that she is up-to-date after a “very long time”?

$$U_i = \{\text{Alice being up-to-date in week } i\} \quad B_i = \{\text{Alice being behing in week } i\}$$

⁶In other words $\{A_j\}_{j \in \{1, 2, \dots, n\}}$ is a partition of Ω of such that $\mathbf{P}(A_j) > 0$ for all $j \in \{1, 2, \dots, n\}$

Furthermore, we know that

$$\begin{array}{lcl} \mathbf{P}(U_{i+1}|U_i) = 0.8 & & \mathbf{P}(U_{i+1}|B_i) = 0.4 \\ \mathbf{P}(B_{i+1}|U_i) = 0.2 & \text{and} & \mathbf{P}(B_{i+1}|B_i) = 0.6 \end{array}$$

Note that $U_i = B_i^c$ for all i by definition, i.e. $\{U_i, B_i\}$ is a partition of the sample space and as a result of total probability theorem we know that

$$\mathbf{P}(U_{i+1}) = \mathbf{P}(U_i) \mathbf{P}(U_{i+1}|\mathbf{P}(U_i)) + \mathbf{P}(B_i) \mathbf{P}(U_{i+1}|\mathbf{P}(B_i)) = 0.8\mathbf{P}(U_i) + 0.4\mathbf{P}(B_i) \quad (3.5)$$

$$\mathbf{P}(B_{i+1}) = \mathbf{P}(U_i) \mathbf{P}(B_{i+1}|\mathbf{P}(U_i)) + \mathbf{P}(B_i) \mathbf{P}(B_{i+1}|\mathbf{P}(B_i)) = 0.2\mathbf{P}(U_i) + 0.6\mathbf{P}(B_i) \quad (3.6)$$

Thus

$$\begin{bmatrix} \mathbf{P}(U_{i+1}) \\ \mathbf{P}(B_{i+1}) \end{bmatrix} = \begin{bmatrix} 0.8 & 0.4 \\ 0.2 & 0.6 \end{bmatrix} \begin{bmatrix} \mathbf{P}(U_i) \\ \mathbf{P}(B_i) \end{bmatrix} \Rightarrow \begin{bmatrix} \mathbf{P}(U_i) \\ \mathbf{P}(B_i) \end{bmatrix} = \left(\begin{bmatrix} 0.8 & 0.4 \\ 0.2 & 0.6 \end{bmatrix} \right)^i \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

One can calculate i^{th} power of $\begin{bmatrix} 0.8 & 0.4 \\ 0.2 & 0.6 \end{bmatrix}$ iteratively (or by using the eigen values and vectors).

However, in addition to (3.5) and (3.6) we also have $\mathbf{P}(U_i) = 1 - \mathbf{P}(B_i)$. Thus we get

$$\begin{aligned} \mathbf{P}(U_{i+1}) &= 0.4 + 0.4 \cdot \mathbf{P}(U_i) &&= 0.4 + 0.4^2 + \dots + 0.4^{i-1} + 0.4^i \mathbf{P}(U_1) \\ &= 0.4 \cdot \frac{1 - 0.4^{i-1}}{0.6} + 0.8 \cdot 0.4^i &&= \frac{2}{3} + \frac{1}{3} \cdot 0.4^i \end{aligned}$$

3.5 Bayes' Rule

Let A_1, A_2, \dots, A_n be disjoint events satisfying⁷ both $\mathbf{P}(A_j) > 0$ for all $j \in \{1, 2, \dots, n\}$ and $\bigcup_{i=1}^n A_j = \Omega$. Then for any event with B such that $\mathbf{P}(B) > 0$, we have

$$\mathbf{P}(A_i|B) = \frac{\mathbf{P}(A_i) \mathbf{P}(B|A_i)}{\mathbf{P}(B)} \quad (3.7)$$

$$= \frac{\mathbf{P}(A_i) \mathbf{P}(B|A_i)}{\sum_{j=1}^n \mathbf{P}(A_j) \mathbf{P}(B|A_j)}. \quad (3.8)$$

Why?

$$\begin{aligned} \mathbf{P}(A_i|B) &= \frac{\mathbf{P}(A_i \cap B)}{\mathbf{P}(B)} && \text{by the definition of conditional probability} \\ &= \frac{\mathbf{P}(A_i) \mathbf{P}(B|A_i)}{\mathbf{P}(B)} && \text{by the definition of conditional probability} \\ &= \frac{\mathbf{P}(A_i) \mathbf{P}(B|A_i)}{\sum_{j=1}^n \mathbf{P}(A_j) \mathbf{P}(B|A_j)}. && \text{by the total probability theorem} \end{aligned}$$

A_i : Possible cause of an effect

B: Effect

$\mathbf{P}(A_i)$: Prior probability

$\mathbf{P}(A_i|B)$: Posterior probability

⁷In other words $\{A_j\}_{j \in \{1, 2, \dots, n\}}$ is a partition of Ω of such that $\mathbf{P}(A_j) > 0$ for all $j \in \{1, 2, \dots, n\}$

Example 3.9. Recall Example 3.7. A fair tetrahedral die rolled once. If the result is either 1 or 2, then it is rolled once more, otherwise there is no second roll. Given that the sum of all the rolls is greater than or equal to 4, what is the probability that outcome of the first roll is i ?

For $A_i = \{\text{The result of the first roll is } i\}$ and $B = \{\text{The sum is as least 4}\}$, we have

$$\begin{aligned} \mathbf{P}(A_i) &= \frac{1}{4} & \forall i \in \{1, 2, 3, 4\} \\ \mathbf{P}(B|A_1) &= \frac{2}{4} & \mathbf{P}(B|A_2) &= \frac{3}{4} & \mathbf{P}(B|A_3) &= 0 & \mathbf{P}(B|A_4) &= 1 \\ \mathbf{P}(B) &= \frac{9}{16} \\ \mathbf{P}(A_1|B) &= \frac{2}{9} & \mathbf{P}(A_2|B) &= \frac{3}{9} & \mathbf{P}(A_3|B) &= 0 & \mathbf{P}(A_4|B) &= \frac{4}{9} \end{aligned}$$

Example 3.10. One in every thousand individuals are affected by a certain rare disease. A test for the disease is correct with probability 0.95 both for individuals with the disease and without the disease. An individual chosen randomly gets a positive test result. What is the probability that she has the disease?

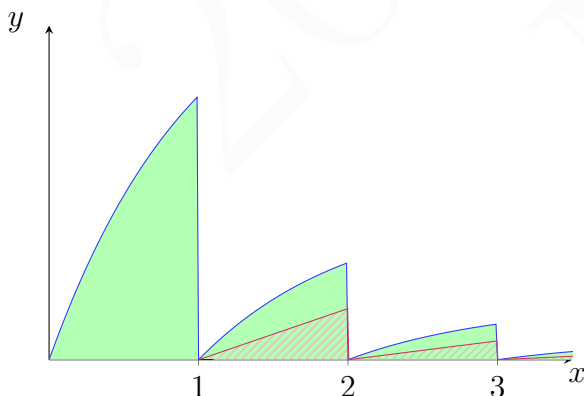
$$\begin{aligned} A &= \{\text{Person has the disease}\} & B &= \{\text{Test is positive}\} \\ \mathbf{P}(A) &= 0.001 \\ \mathbf{P}(B|A) &= 0.95 \\ \mathbf{P}(B|A^c) &= 0.05 \\ \mathbf{P}(A|B) &= \frac{0.001 \cdot 0.95}{0.001 \cdot 0.95 + 0.999 \cdot 0.05} \approx 0.01866 \end{aligned}$$

Example 3.11 (The Monty Hall Problem).

Example 3.12. Consider the probability model with the continuous uniform probability law on the samples space $\Omega = \{(x, y) : x \geq 0, 0 < y \leq e^{-\lfloor x \rfloor} - e^{-x}\}$. Thus $\exists \beta > 0$ such that

$$\begin{aligned} \mathbf{P}(A) &= \beta \cdot \text{Area of } A. \\ A_j &= \{(x, y) \in \Omega : x \in [j, j + 1)\} & \forall j \in \mathbb{Z}_{\geq 0} \\ B &= \{(x, y) \in \Omega : x \geq 1 \text{ and } y < \frac{x - \lfloor x \rfloor}{3} e^{-\lfloor x \rfloor}\} \end{aligned}$$

- (a) Determine the value of β
- (b) What is the probability of the event B ?
- (c) What is the conditional probability of the event A_j given event B ?



$$\begin{aligned} \mathbf{P}(A_j) &= \beta e^{-j} \int_0^1 (1 - e^{-t}) dt \\ &= \beta e^{-(j+1)} & \forall j \in \mathbb{Z}_{\geq 0} \end{aligned}$$

A_j 's form a partition of Ω . Then as a result of the countable additivity of the probability

$$\begin{aligned} \mathbf{P}(\Omega) &= \sum_{j=0}^{\infty} \mathbf{P}(A_j) = \sum_{j=0}^{\infty} \beta e^{-(j+1)} \\ &= \beta \frac{e^{-1}}{1 - e^{-1}} = \beta \frac{1}{e - 1} \end{aligned}$$

Thus $\beta = e - 1$.

In order to calculate $\mathbf{P}(\mathbf{B})$ we can use the total probability theorem

$$\begin{aligned}\mathbf{P}(\mathbf{B}|\mathbf{A}_0) &= 0 \\ \mathbf{P}(\mathbf{B}|\mathbf{A}_j) &= \frac{\frac{1}{2} \cdot 1 \cdot \frac{e^{-j}}{3}}{e^{-j} \int_0^1 (1 - e^{-t}) dt} = \frac{e}{6} \quad \forall j \in \mathbb{Z}_+ \\ \mathbf{P}(\mathbf{B}) &= \sum_{j=0}^{\infty} \mathbf{P}(\mathbf{A}_j) \mathbf{P}(\mathbf{B}|\mathbf{A}_j) = \frac{e-1}{6} \sum_{j=1}^{\infty} e^{-j} \\ &= \frac{1}{6}\end{aligned}$$

We can calculate $\mathbf{P}(\mathbf{A}_j|\mathbf{B})$'s using Bayes' rule

$$\mathbf{P}(\mathbf{A}_j|\mathbf{B}) = \frac{\mathbf{P}(\mathbf{A}_j) \mathbf{P}(\mathbf{B}|\mathbf{A}_j)}{\mathbf{P}(\mathbf{B})} = \begin{cases} 0 & j = 0 \\ (e-1)e^{-j} & j \in \mathbb{Z}_+ \end{cases}$$

4 Independence

4.1 Independence Of Two Events

Definition 4.1. Two events A and B are independent iff

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

Example 4.1. Consider two successive coin tosses in which all four outcomes are equally likely, i.e., $\Omega = \{HH, HT, TH, TT\}$ and $\mathbf{P}(\cdot)$ is the discrete uniform probability law.

$H_i = \{i^{\text{th}} \text{ coin toss results in a head}\},$

$G = \{\text{at least one of the coin tosses is a head}\}.$

$D = \{\text{results of the two coin tosses are different}\}, \quad S = \{\text{results of the two coin tosses are same}\},$

Are the events H_1 and H_2 independent? Is the event D independent of the event $H_1/S/G$?

$$H_1 = \{HH, HT\}$$

$$H_2 = \{HH, TH\}$$

$$D = \{HT, TH\}$$

$$S = \{HH, TT\}$$

$$G = \{HT, TH, HH\}$$

$$\mathbf{P}(H_1) = 1/2$$

$$\mathbf{P}(H_2) = 1/2$$

$$\mathbf{P}(D) = 1/2$$

$$\mathbf{P}(S) = 1/2$$

$$\mathbf{P}(G) = 3/4$$

Then

$$H_1 \cap H_2 = \{HH\}$$

$$D \cap H_1 = \{HT\}$$

$$D \cap S = \emptyset$$

$$D \cap G = \{HT, TH\}$$

$$\mathbf{P}(H_1 \cap H_2) = 1/4$$

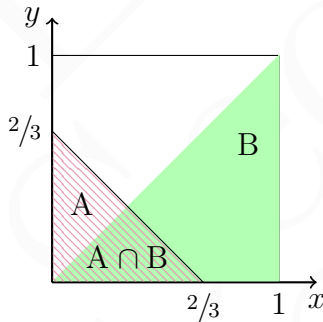
$$\mathbf{P}(D \cap H_1) = 1/4$$

$$\mathbf{P}(D \cap S) = 0$$

$$\mathbf{P}(D \cap G) = 1/2$$

H_1 and H_2 are independent. D and H_1 are independent. D and S are not independent. D and G are not independent.

Example 4.2. Recall the continuous uniform probability law considered in Example 2.6



$$\Omega = \{(x, y) : x \in (0, 1) \text{ and } y \in (0, 1)\}$$

$$\mathbf{P}(E) = \text{the area of the event } E \quad \forall E \in \mathcal{F}$$

$$A = \{(x, y) \in \Omega : x + y < 2/3\}$$

$$B = \{(x, y) \in \Omega : x > y\}$$

$$\mathbf{P}(A) = 2/9$$

$$\mathbf{P}(B) = 1/2$$

$$\mathbf{P}(A \cap B) = 1/9$$

Remark 4.1.

(i) If A and B are independent events and $\mathbf{P}(A) > 0$ then $\mathbf{P}(B|A) = \mathbf{P}(B)$.

(ii) If A and B are independent then A^c and B are independent.

$$\begin{aligned} \mathbf{P}(A^c \cap B) &= \mathbf{P}(B) - \mathbf{P}(A \cap B) && \text{by additivity because } (A^c \cap B) \cap (A \cap B) = \emptyset \\ &= \mathbf{P}(B) - \mathbf{P}(A) \mathbf{P}(B) && \text{because A and B are independent.} \\ &= \mathbf{P}(B) \mathbf{P}(A^c) \end{aligned}$$

(iii) If $\mathbf{P}(A) = 0$ then events A and B are independent for any event B.

(iv) If $\mathbf{P}(A) = 1$ then events A and B are independent for any event B.

(v) If events A and B are independent and $A \cap B = \emptyset$ then either $\mathbf{P}(A) = 0$ or $\mathbf{P}(B) = 0$.

“If two independent events are mutually exclusive then at least one of them has zero probability.”

(vi) If events A and B are independent and $A \subset B$ then either $\mathbf{P}(A) = 0$ or $\mathbf{P}(B) = 1$.

4.2 Conditional Independence

Definition 4.2. Let C be an event with a positive probability, i.e., $\mathbf{P}(C) > 0$. Then events A and B are conditionally independent given C iff

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

Recall as a result of multiplication rule

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

If A and B are conditionally independent given C and $\mathbf{P}(A \cap C) > 0$, then $\mathbf{P}(B|C) = \mathbf{P}(B|A \cap C)$.

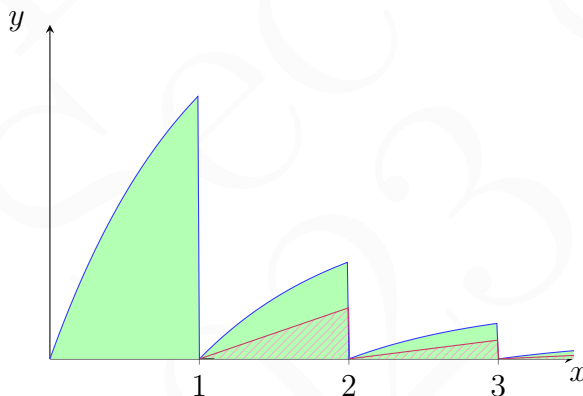
Example 4.3 (Independence $\not\Rightarrow$ Conditional Independence). Recall Example 4.1. Events H_1 and H_2 are independent. But H_1 and H_2 not conditionally independent given event D because

$$\begin{aligned} H_1 \cap D = \{HT\} &\Rightarrow \mathbf{P}(H_1 \cap D) = 1/2 &\Rightarrow \mathbf{P}(H_1|D) = 1/2 \\ H_2 \cap D = \{TH\} &\Rightarrow \mathbf{P}(H_2 \cap D) = 1/4 &\Rightarrow \mathbf{P}(H_2|D) = 1/2 \\ H_1 \cap H_2 \cap D = \{HT\} &\Rightarrow \mathbf{P}(H_1 \cap H_2 \cap D) = 0 &\Rightarrow \mathbf{P}(H_1 \cap H_2|D) = 0 \neq 1/2 \cdot 1/2. \end{aligned}$$

H_1 and H_2 not conditionally independent given event G :

$$\mathbf{P}(H_1|G) = 2/3 \quad \mathbf{P}(H_2|G) = 2/3 \quad \mathbf{P}(H_1 \cap H_2|G) = 1/3 \neq 2/3 \cdot 2/3.$$

Example 4.4 (Independence \neq Conditional Independence). Recall the probability model with the samples space $\Omega = \{(x, y) : x \geq 0, 0 < y \leq e^{-\lfloor x \rfloor} - e^{-x}\}$ and the continuous uniform probability law considered in Example 3.12.



$$\begin{aligned} \mathbf{P}(A) &= (e - 1) \cdot \text{Area of } A. \\ A_j &= \{x \in [j, j + 1)\} \quad \forall j \in \mathbb{Z}_{\geq 0} \\ B &= \{x \geq 1 \text{ and } y < \frac{x - \lfloor x \rfloor}{3} e^{-\lfloor x \rfloor}\} \end{aligned}$$

Recall that

$$\begin{aligned} \mathbf{P}(A_j) &= (e - 1)e^{-(j+1)} \quad \forall j \in \mathbb{Z}_{\geq 0} \\ \mathbf{P}(A_j|B) &= \begin{cases} 0 & j = 0 \\ (e - 1)e^{-j} & j \in \mathbb{Z}_+ \end{cases} \end{aligned}$$

Thus for any non-negative integer j the events A_j and B are not independent.

Let k be a positive integer and the event C be

$$\begin{aligned} C &= \{(x, y) \in \Omega : x \geq 2\} &= \bigcup_{j=2}^{\infty} A_j \\ \mathbf{P}(C) &= e^{-2} \end{aligned}$$

For any non-negative integer j the events A_j and B are conditionally independent given C because

$$\begin{aligned} \mathbf{P}(A_j|C) &= \begin{cases} 0 & j < k \\ (e - 1)e^{1-j} & j \geq k \end{cases} \\ \mathbf{P}(A_j|C \cap B) &= \begin{cases} 0 & j < k \\ (e - 1)e^{1-j} & j \geq k \end{cases} \end{aligned}$$

Example 4.5. In certain situations conditional independence is a more reasonable assumption than the independence. Recall, for instance, the framework of Example 3.10 in which a rare disease affecting one in every thousand individuals was considered. A test for the disease which is correct with probability 0.95 both for individuals with the disease and without the disease, was considered. Let us assume that events of getting a positive result in the first test and in the second test are conditionally independent given that the individual has the disease and also given that the individual is disease free. Calculate the probability of an individual with two positive test results to have the disease.

$$A = \{\text{Person has the disease}\} \quad B_i = \{i^{\text{th}} \text{ test is positive}\}$$

$$\begin{aligned} \mathbf{P}(B_2 \cap B_1|A) &= \mathbf{P}(B_2|A) \mathbf{P}(B_1|A) \\ &= (0.95)^2 \end{aligned} \quad (4.1)$$

$$\begin{aligned} \mathbf{P}(B_2 \cap B_1|A^c) &= \mathbf{P}(B_2|A^c) \mathbf{P}(B_1|A^c) \\ &= (0.05)^2 \end{aligned} \quad (4.2)$$

$$\begin{aligned} \mathbf{P}(A|B_1 \cap B_2) &= \frac{0.001 \cdot (0.95)^2}{0.001 \cdot (0.95)^2 + 0.999 \cdot (0.05)^2} & \mathbf{P}(A|B_1) &= \frac{0.001 \cdot 0.95}{0.001 \cdot 0.95 + 0.999 \cdot 0.05} \\ &\approx 0.2654 & &\approx 0.01866 \end{aligned}$$

Remark 4.2. It is important to note that the conditional independence is merely an assumption in the above example. The reason for a test to fail to detect the disease for an individual with the disease might depend on a feature of the test or a trait present in a certain fraction of the population, in either case the conditional independence assumption described in (4.1) will not be true. Similarly the reasons for a test to give a false positive result for disease free individuals can be caused by a feature of the test or a trait present in a certain fraction of the population, in either case the conditional independence assumption described in (4.2) will not be true.

4.3 Independence of A Collection of Events

Definition 4.3. Events A_1, A_2, \dots, A_n are independent iff

$$\mathbf{P}\left(\bigcap_{j \in S} A_j\right) = \prod_{j \in S} \mathbf{P}(A_j) \quad \forall S \subset \{1, 2, \dots, n\}. \quad (4.3)$$

Events A_1, A_2, \dots, A_n are pairwise independent iff

$$\mathbf{P}\left(A_i \cap A_j\right) = \mathbf{P}(A_i) \mathbf{P}(A_j) \quad \forall j \neq i. \quad (4.4)$$

The independence of a collection of events, is also called mutual independence.

If A_1, A_2, \dots, A_n are independent then $\{A_j\}_{j \in J}$ are independent for any $J \subset \{1, \dots, n\}$.

Note that independence implies pairwise independence.

Example 4.6 (Pairwise independence $\not\Rightarrow$ Independence). Recall two coin tosses with discrete uniform probability law considered in Example 4.1.

$$\begin{array}{lll} H_1 = \{HH, HT\} & H_2 = \{HH, TH\} & D = \{HT, TH\} \\ \mathbf{P}(H_1) = 1/2 & \mathbf{P}(H_2) = 1/2 & \mathbf{P}(D) = 1/2 \\ H_1 \cap H_2 = \{HH\} & H_2 \cap D = \{TH\} & D \cap H_1 = \{HT\} \\ \mathbf{P}(H_1 \cap H_2) = 1/4 & \mathbf{P}(H_2 \cap D) = 1/4 & \mathbf{P}(D \cap H_1) = 1/4 \end{array}$$

H_1, H_2, D are pairwise independent.

$$H_1 \cap H_2 \cap D = \emptyset \quad \mathbf{P}(H_1 \cap H_2 \cap D) = 0$$

H_1, H_2, D are not independent.

Example 4.7. Consider two rolls a six faced die with the discrete uniform probability law,

$$\begin{aligned} \Omega &= \{(x, y) : x \in \{1, 2, 3, 4, 5, 6\} \text{ and } y \in \{1, 2, 3, 4, 5, 6\}\} \\ \mathbf{P}(E) &= |E|/36 && \forall E \in \mathcal{2}^\Omega \\ A &= \{(x, y) \in \Omega : x \in \{1, 2, 3\}\} \\ B &= \{(x, y) \in \Omega : x \in \{3, 4, 5\}\} \\ C &= \{(x, y) \in \Omega : x + y = 9\} && = \{(3, 6), (4, 5), (5, 4), (6, 3)\} \end{aligned}$$

Then $\mathbf{P}(A \cap B \cap C) = \mathbf{P}(A) \cdot \mathbf{P}(B) \cdot \mathbf{P}(C)$, in particular

$$\mathbf{P}(A) = 1/2 \qquad \mathbf{P}(B) = 1/2 \qquad \mathbf{P}(C) = 1/9 \qquad \mathbf{P}(A \cap B \cap C) = 1/36.$$

But

$$\begin{aligned} \mathbf{P}(A \cap B) &= 1/6 && \neq \mathbf{P}(A) \mathbf{P}(B) \\ \mathbf{P}(A \cap C) &= 1/36 && \neq \mathbf{P}(A) \mathbf{P}(C) \\ \mathbf{P}(B \cap C) &= 1/9 && \neq \mathbf{P}(B) \mathbf{P}(C) \end{aligned}$$

HW. For Example 4.5 assume that B_1, B_2, B_3 are conditionally independent given A and B_1, B_2, B_3 are conditionally independent given A^c . What is the probability that an individual with three positive tests has the disease? (≈ 0.8729)

One might ask why the condition (4.3) is imposed for all subset of $\{1, \dots, n\}$ for independence. There is no simple answer. But understanding the implications of that definition will help us appreciate that choice.

Recall that A and B are independent iff A^c and B are independent. There is a counter part to that for a collection of events: Events $\{A_j\}_{j \in \{1, \dots, n\}}$ are independent iff for any $J \subset \{1, \dots, n\}$, the events $\{A_j^c\}_{j \in J} \cup \{A_j\}_{j \in \{1, \dots, n\} \setminus J}$ are independent. (iff for any $J \subset \{1, \dots, n\}$ if we replace A_j 's for $j \in J$ with their complements we will get an independent collection of events.)

HW. Using (4.3) prove that A_1, A_2, \dots, A_n are independent iff A_1^c, A_2, \dots, A_n are independent.

If A_1, A_2, A_3 are independent then A_1 and $A_2 \cap A_3$ is independent. This follows from (4.3). On the other hand if A_1, A_2, A_3 are independent then A_1, A_2^c, A_3^c because of the property we have discussed in the preceding paragraph. Hence A_1 and $A_2^c \cap A_3^c$ are independent. Then A_1 and $(A_2^c \cap A_3^c)^c$ (i.e., $A_2 \cup A_3$) are independent. If A_1, \dots, A_n are independent then using similar considerations we can conclude that any event that can be expressed in terms of union, intersections, and complementations of the events A_2, \dots, A_n is independent of A_1 .

Above feature of independence provides us a characterization that is more abstract but in a sense more complete and satisfactory. A finite collection of events A_1, A_2, \dots, A_n are independent iff for any partition $\{J_i\}_{i \in I}$ of the index set $\{1, 2, \dots, n\}$, any choice of events $\{B_i\}_{i \in I}$ such that each B_i that can be expressed in terms of union, intersections, and complementations of the events $\{A_j\}_{j \in J_i}$, we have

$$\mathbf{P}\left(\bigcap_{i \in I} B_i\right) = \prod_{i \in I} \mathbf{P}(B_i). \tag{4.5}$$

For a finite collection of events A_1, A_2, \dots, A_n , the set of all events that can be expressed in terms of union, intersections, and complementations of the events A_1, A_2, \dots, A_n is denoted by $\sigma(A_1, A_2, \dots, A_n)$.

HW (*Only if you are really curious about $\sigma(\cdot)$*). For any finite collection of events A_1, A_2, \dots, A_n there is a partition of $\{D_i\}_{i \in I}$ of Ω satisfying both $|I| \leq 2^n$ and $A_j = \bigcup_{i \in I_j} D_i$ for some $I_j \subset I$ for all $j \in \{1, \dots, n\}$. Furthermore, $|\sigma(A_1, A_2, \dots, A_n)| = 2^{|I|}$.

4.4 Independent Trials

A random experiment that is composed of independent and identical stages is called independent trials. We have already seen instances of independent trials before when we considered two rolls of a four faced die in Example 2.3, and a six faced die in Example 4.7, and two tosses of a fair coin in Example 2.6. Even the random experiment discussed via a continuous model in Example 2.6 is an independent trial.

In an independent trial one can specify the probability law of the overall random experiment by describing the probability law of a single trial because of the independence and identicalness. We have not used this fact in our discussions before because we have not introduced the concept of independence then. Before stating this relation more generally let us discuss a very instructive special case independent Bernoulli Trials

Example 4.8 (Independent Bernoulli Trials and Binomial Probabilities). Consider the random experiment of n -coin flips, with the sample space

$$\Omega = \{(x_1, x_2, \dots, x_n) : x_j \in \{H, T\} \forall j \in J\} \quad \text{where} \quad J = \{1, 2, \dots, n\}.$$

Let H_j be the event of getting an H in j^{th} coin toss, i.e.,

$$H_j = \{(x_1, x_2, \dots, x_n) \in \Omega : x_j = H\}.$$

Thus H_j^c , i.e., the complement of H_j , is the event of getting a T in j^{th} coin toss.

Let us assume the probability of getting an H in j^{th} coin toss is p , i.e.,

$$\mathbf{P}(H_j) = p.$$

Events about different trials are independent because we have independent trials. Thus events $\{H_j\}_{j \in J}$ are independent. More generally for any $I \subset J$ events $\{H_j\}_{j \in I} \cup \{H_j^c\}_{j \in J \setminus I}$ are independent as well. Thus

$$\begin{aligned} \mathbf{P}\left(\left(\bigcap_{j \in I} H_j\right) \cap \left(\bigcap_{j \in J \setminus I} H_j^c\right)\right) &= \left(\prod_{j \in I} \mathbf{P}(H_j)\right) \cdot \left(\prod_{j \in J \setminus I} \mathbf{P}(H_j^c)\right) \\ &= p^{|I|} (1-p)^{|J \setminus I|} \\ &= p^{|I|} (1-p)^{n-|I|} \end{aligned}$$

Note that $\left(\bigcap_{j \in I} H_j\right) \cap \left(\bigcap_{j \in J \setminus I} H_j^c\right)$ is a singleton; it is the string whose entries for indices in I are H and whose indices for $J \setminus I$ is T . Thus we have determined the probability of all of the singletons and thus the probability law of the independent Bernoulli trials.

Let the event A_k for $k \in \{0, 1, \dots, n\}$ be having k heads and $n - k$ tails in n trials:

$$A_k = \{(x_1, \dots, x_n) \in \Omega : x_j = H \forall j \in I \text{ and } x_j = T \forall j \notin I \text{ for some size } k \text{ set } I\}$$

Note that there are $\binom{n}{k}$ outcomes in A_k and corresponding singletons each happen with probability $p^k(1-p)^{n-k}$. Thus

$$\mathbf{P}(A_k) = \binom{n}{k} p^k (1-p)^{n-k} \quad (4.6)$$

$\mathbf{P}(A_k)$ are called binomial probabilities and the additivity of probability and $\Omega = \cup_{k=0}^n A_k$ imply the following equality, called binomial formula.

$$\sum_{k=0}^n \binom{n}{k} p^k (1-p)^{n-k} = 1.$$

More on binomial coefficient $\binom{n}{k}$ in the following section.

For an (identical) independent trial, the sample space $\Omega^{(n)}$ will be the Cartesian product of sample spaces for individual trials, $\Omega_1, \dots, \Omega_n$, and Ω_j 's will be identical, recall Bernoulli trials.

$$\begin{aligned} \Omega^{(n)} &= \Omega_1 \times \dots \times \Omega_n \\ &= \{(\omega_1, \dots, \omega_n) : \omega_j \in \Omega_j \forall j \in \{1, 2, \dots, n\}\} \end{aligned}$$

The set of all events for independent trials $\mathcal{F}^{(n)}$ will be determined by the set of all events of individual trials $\mathcal{F}_1, \dots, \mathcal{F}_n$. For all $\widetilde{A}_j \in \mathcal{F}_j$ we can define the corresponding event for the independent trials as

$$A_j = \{(\omega_1, \dots, \omega_n) \in \Omega^{(n)} : \omega_j \in \widetilde{A}_j\} \quad \forall \widetilde{A}_j \in \mathcal{F}_j. \tag{4.7}$$

The set of all events $\mathcal{F}^{(n)}$ should include all events of the form given above and it is smallest such σ -algebra. This condition ensures that each event we have for individual trials will have a corresponding event in the probability space for independent trial. Note for instance having a head in the j^{th} trial, H_j , is an event for the probability space we have for independent Bernoulli trials.

Let us denote the probability law associated with j^{th} trial by $\mathbf{P}_j(\cdot)$. For the probability law $\mathbf{P}(\cdot)$ of the (identical) independent trials to be consistent with the probability laws we have for individual trials the following condition has to be satisfied

$$\mathbf{P}(A_j) = \mathbf{P}_j(\widetilde{A}_j) \quad \forall j \in \{1, \dots, n\} \text{ and } \forall \widetilde{A}_j \in \mathcal{F}_j. \tag{4.8}$$

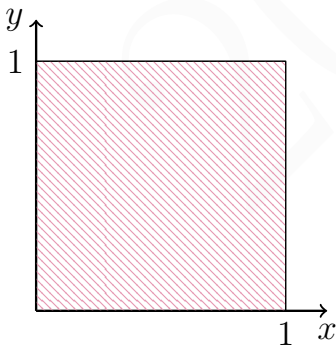
In the case of Bernoulli trials this is essentially setting $\mathbf{P}(H_j) = p$.

If one assumes the independence of the events A_1, A_2, \dots, A_n of the form given in (4.7), then (4.8) uniquely determines the probability law for independent trials. We can restate (4.8) together with the independence as follows.

$$\mathbf{P}\left(\bigcap_{j \in J} A_j\right) = \prod_{j \in J} \mathbf{P}_j(\widetilde{A}_j) \quad \forall J \subset \{1, \dots, n\} \text{ and } \widetilde{A}_j \in \mathcal{F}_j. \tag{4.9}$$

This observation is similar to obtaining the probability law for the independent Bernoulli trials just by assuming the independence of H_1, \dots, H_n and by setting $\mathbf{P}(H_j) = p$. For countable sample spaces this is plausible because in most case we can determine the probabilities of the singletons, as we did for independent Bernoulli trials. For uncountable sample spaces picture is more nuanced. But often the intuition one gets from the following example is helpful.

Example 4.9 (Example 2.6).



$$\begin{aligned} \Omega &= \{(x, y) : x \in (0, 1) \text{ and } y \in (0, 1)\} \\ \mathbf{P}(A) &= \text{the area of the event } A & \forall A \in \mathcal{F} \\ \Omega_j &= \{x : x \in (0, 1)\} & \forall j \in \{1, 2\} \\ \mathbf{P}_j(\widetilde{A}_j) &= \text{the length of the event } \widetilde{A}_j & \forall j \in \{1, 2\}, \forall \widetilde{A}_j \in \mathcal{F}_j \end{aligned}$$

5 Counting

In principle, counting is trivial when a list of elements can be provided. More often than not, however, either counting has to be done indirectly, or counting can be done much more efficiently when done indirectly. In the following, we will discuss some of the most frequently used concepts and methods for counting.

The Counting Principle.

- A processes with $r \in \mathbb{Z}_+$ stages.
- There are n_j possible results at stage j for each $j \in \{1, \dots, r\}$
- The # of possible results is $n_1 \cdot n_2 \cdots n_r$

The number of choices at each stage needs to be the same but the actual set these choices come from can change.

Example 5.1. How many different license plates can be formed with 2 letters from Latin alphabet (used for English) and 3 digits?

$$26 \cdot 26 \cdot 10 \cdot 10 \cdot 10$$

How about when letter or number (arithmetic numeral) can be repeated?

$$26 \cdot 25 \cdot 10 \cdot 9 \cdot 8$$

Permutations.

Any ordering of n distinct objects is called a permutation.

A set of n distinct objects has $n \cdot (n - 1) \cdots 1$ different permutations.

Any length k ordering of elements of a set is called a k -permutation.

A set of n distinct objects has $n \cdot (n - 1) \cdots (n - k + 1)$ different k -permutations $\forall k \leq n$.

n factorial, denoted by $n!$, is defined for any non-negative integer n as follows,

$$n! = \begin{cases} n \cdot (n - 1) \cdot (n - 2) \cdots 1 & n \in \mathbb{Z}_+ \\ 1 & n = 0 \end{cases} \quad (5.1)$$

- The # of permutations for n objects is $n!$
- The # of k -permutations for n objects is $\frac{n!}{(n-k)!}$

Example 5.2. How many ways can you order

- n_1 classical music CDs
- n_2 rock music CDs
- n_3 country music CDs

on a shelf so that CDs of the same genre are contiguous?

$$3! \cdot n_1! \cdot n_2! \cdot n_3!$$

How many ways can you order k_1 of the n_1 classical music CDs k_2 of the n_2 rock music CDs k_3 of the n_3 country music CDs on a shelf so that CDs of the same genre are contiguous?

$$3! \cdot \frac{n_1!}{(n_1 - k_1)!} \cdot \frac{n_2!}{(n_2 - k_2)!} \cdot \frac{n_3!}{(n_3 - k_3)!}$$

For any non-negative integer n and any non-negative integer $k \leq n$, the binomial coefficient $\binom{n}{k}$ is defined as

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{5.2}$$

Combinations.

Any size k subset of n distinct objects is called a size k combination.

Any n distinct objects has $\binom{n}{k}$ distinct size k combinations because every size k combination will correspond to $k!$ distinct size k permutations.

- $\binom{n}{n} = 1$: Consistent with the observation that the only size n subset of size n set S is S .
- $\binom{n}{0} = 1$: Consistent with the observation that the only size 0 subset of size n set S is \emptyset .
- $\sum_{k=0}^n \binom{n}{k} = ?$ How many subsets does a size n set S has?

$$\sum_{k=0}^n \binom{n}{k} = 2^n$$

Example 5.3. For 10 independent tosses of a coin which results in a H with probability p . What is the conditional probability that first two coin tosses are H given that only 3 of the coin tosses are H ?

$A = \{\text{The first two coin tosses are } H\}$

$B = \{\text{3 out of 10 coin tosses are } H\}$

$$\mathbf{P}(A|B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)} = \frac{|A \cap B| p^3 (1-p)^7}{|B| p^3 (1-p)^7} = \frac{\binom{8}{1}}{\binom{10}{3}} = \frac{8}{\frac{10 \cdot 9 \cdot 8}{3 \cdot 2}} = \frac{1}{15}$$

For any positive integer r and r -tuple of non-negative integers (n_1, \dots, n_r) satisfying $\sum_{j=1}^r n_j = n$ the multinomial coefficient $\binom{n}{n_1, \dots, n_r}$ is defined as

$$\binom{n}{n_1, \dots, n_r} = \frac{n!}{n_1! \cdots n_r!} \tag{5.3}$$

Partitions Of a Given Size Vector/Type (n_1, \dots, n_r) . n objects can be divided into r groups with j^{th} group having size n_j in $\binom{n}{n_1, \dots, n_r}$ different ways because every such partition corresponds to $n_1! \cdots n_r!$ distinct permutations.

Example 5.4 (Multinomial Probabilities). Consider an r -faced die for which j^{th} face is up with probability p_j . For n independent rolls of the die what is the probability that n_j of the rolls are j for all $j \in \{1, \dots, r\}$ where $n_1 + \dots + n_r = n$.

$$\Omega = \{(x_1, \dots, x_n) : x_i \in \{1, \dots, r\} \forall i \in \{1, \dots, n\}\}$$

$$\mathbf{P}(\{\omega\}) = \prod_{j=1}^r (p_j)^{n_j} \quad \forall \omega \in \Omega : n_j \text{ of the rolls are } j$$

$$= (p_1)^{n_1} \cdots (p_r)^{n_r}$$

$$\mathbf{P}(A_{(n_1, \dots, n_r)}) = \binom{n}{n_1, \dots, n_r} \prod_{j=1}^r (p_j)^{n_j}$$

$$= \frac{n!}{n_1! \cdots n_r!} (p_1)^{n_1} \cdots (p_r)^{n_r}$$

Example 5.5. How many anagrams do the word TATTOO have?

1	2	3	4	5	6
---	---	---	---	---	---

 We will decide

- Which one of these 6 positions will be assigned A ?
- Which two of these 6 positions will be assigned O ?
- Which three of these 6 positions will be assigned T ?

There are $\binom{6}{3,2,1}$ ways to do it.

$$\binom{6}{3,2,1} = \frac{6!}{3! \cdot 2! \cdot 1!} = \frac{6 \cdot 5 \cdot 4}{2} = 60.$$

Example 5.6. 4 graduate and 12 under graduate students are to form 4 groups each with 4 students. If all assignment of 16 students to 4 groups are have equal probability what is the probability that each group will have exactly one graduate student.

$$\mathbf{P}(A) = \frac{\binom{4}{1,1,1,1} \binom{12}{3,3,3,3}}{\binom{16}{4,4,4,4}} = 4! \cdot \frac{12! (4!)^4}{(3!)^4 16!} = 4! \cdot \frac{4 \cdot 4 \cdot 4 \cdot 4}{16 \cdot 15 \cdot 14 \cdot 13} = \frac{64}{5 \cdot 7 \cdot 13}$$

We can also solve this problem using a model based on multiplication rule.

$$\Omega = \left\{ (x_1, x_2, \dots, x_{16}) : x_j \in \{a, b, c, d\} \text{ and } \sum_{j=1}^{16} \mathbb{1}_{\{x_j=\tau\}} = 4 \forall \tau \in \{a, b, c, d\} \right\} \tag{5.4}$$

where

$$\mathbb{1}_{\{x_j=\tau\}} = \begin{cases} 1 & \text{if } x_j = \tau \\ 0 & \text{else} \end{cases}.$$

Uniform probability law on Ω is obtained by assigning x_j 's one by one by choosing without replacement values from the set which initially has 4 of each number. Thus the probability each singleton will be

$$\mathbf{P}(\{\omega\}) = \frac{4}{16} \cdot \frac{3}{15} \cdots \frac{1}{1} = \frac{4! \cdot 4! \cdot 4! \cdot 4!}{16!}$$

We have two ways to apply the sequential model using the multiplication rule. For both of them we assume we have $x_1, x_2, x_3,$ and x_4 to be graduate students.

- (i) For $j \in \{1, 2, 3, 4\}$, let A_j be the event that x_j is assigned to a group without a graduate student.

$$\begin{aligned} \mathbf{P}(A) &= \mathbf{P}(A_1 \cap A_2 \cap A_3 \cap A_4) \\ &= \mathbf{P}(A_1) \cdot \mathbf{P}(A_2|A_1) \cdot \mathbf{P}(A_3|A_1 \cap A_2) \cdot \mathbf{P}(A_4|A_1 \cap A_2 \cap A_3) \\ &= \frac{16}{16} \cdot \frac{12}{15} \cdot \frac{8}{14} \cdot \frac{4}{13} = \frac{64}{5 \cdot 7 \cdot 13} \end{aligned}$$

- (ii) Let $A_{abcd} = \{(x_1, x_2, \dots, x_{16}) \in \Omega : (x_1, x_2, x_3, x_4) = (a, b, c, d)\}$. Then

$$\mathbf{P}(A_{abcd}) = \frac{4}{16} \cdot \frac{4}{15} \cdot \frac{4}{14} \cdot \frac{4}{13}$$

On the other hand

$$\mathbf{P}(A) = 4! \cdot \mathbf{P}(A_{abcd}) = \frac{64}{5 \cdot 7 \cdot 13}$$

HW. Show that for group sizes $(3, 4, 4, 5)$, $\mathbf{P}(A) = \frac{12}{91}$ both using a counting argument and using a description via multiplication rule.

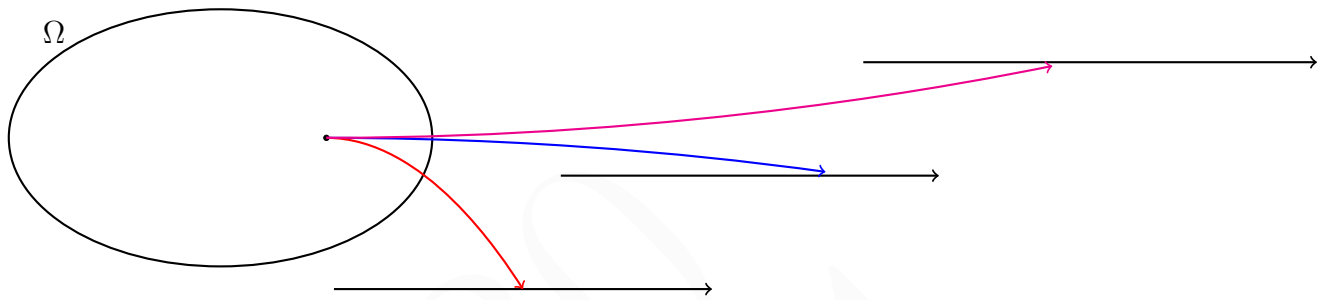
Example 5.7. A type for n objects and r groups is an r -tuple of non-negative integer (n_1, \dots, n_r) satisfying $n_1 + \dots + n_r = n$. How many distinct types are there for n objects and r groups?

There $n+r-1$ items, n of which are objects, and $r-1$ of which are separators. Any ordering of these items is associated with partition as follows:

- All objects before the first separator are assigned to the first group.
- All objects between $(j-1)^{\text{th}}$ and j^{th} separators are assigned to j^{th} group.
- All objects after $(r-1)^{\text{th}}$ separator are assigned to r^{th} group.

Then r -tuple of non-negative integer (n_1, \dots, n_r) satisfying $n_1 + \dots + n_r = n$ is uniquely determined by the places of $r-1$ separators among $n+r-1$ items and vice-versa. Thus there $\binom{n+r-1}{r-1}$ distinct types.

6 Discrete Random Variables



In order to analyze various aspects of random experiments we often use real valued functions on the sample space, called *random variables*.

Example 6.1. Consider the following random experiment:

- Total population of a city, as well as the age, the height, and the weight of its residents is known.
- One resident of the city is chosen uniformly at random among all the residents of the city.

Then the age, the height, and the weight of the resident chosen are all random variables. Note that:

- Functions $A : \Omega \rightarrow \mathbb{Z}_+$, $H : \Omega \rightarrow \mathbb{R}_+$, $W : \Omega \rightarrow \mathbb{R}$ are known before the random experiment is conducted. In other words for all possible outcomes of the experiment (i.e., $\forall \omega \in \Omega$), the value of every single one of these functions (i.e., each one of $A(\omega)$, $H(\omega)$, $W(\omega)$) is known.
- The only thing that is random is the outcome of the experiment (i.e. resident chosen), which determines the values of all the random variables defined on the sample space, i.e. the population (residents) of the city. The probabilistic behavior of this choice is determined by the probability law.

“A random variable is neither random nor variable.”⁸ A random variable is a deterministic function on the sample space but the argument of the function, i.e., the outcome of the random experiment, is random and its behavior is governed by the probability law.

In next few weeks, we will revisit concepts we have discussed before for events such as total probability theorem, Bayes’ rule, independence, and conditional independence for random variables. We will also introduce new concepts for random variables such as expected value and variance.

Before we start our discussion in earnest let us introduce a commonly used notational convention we will adopt for the rest of the course.

- We denote random variables with capital letter: X, Y, Z, U, V .
- We denote the particular real values these random variables get by the corresponding lower case letter: x, y, z, u, v .

Thus in order to say that value of the random variable X is equal to x when the outcome of the random experiment is ω , we write

$$X(\omega) = x$$

⁸This phrase is often attributed to Gian-Carlo Rota

For expressing events about random variables and their probabilities, we use the following shorthand representations

$$\begin{aligned}\{\omega \in \Omega : \mathbf{X}(\omega) \in S\} &= \{\mathbf{X}(\omega) \in S\}, \\ &= \{\mathbf{X} \in S\}, \\ \mathbf{P}(\{\omega \in \Omega : \mathbf{X}(\omega) \in S\}) &= \mathbf{P}(\{\mathbf{X} \in S\}), \\ &= \mathbf{P}(\{\mathbf{X} \in S\}), \\ &= \mathbf{P}(\mathbf{X} \in S).\end{aligned}$$

If consider random variables that can only take countably many distinct values, then certain technical nuances disappear and a precise and rigorous discussion becomes accessible to any one whose is familiar with infinite sums. Thus we will restrict our discussion exclusively for those random variables in next few weeks.

6.1 Probability Mass Function

Definition 6.1. A real valued function \mathbf{X} on the sample space Ω is a discrete random variable iff there exists a countable set Θ satisfying the following two conditions

- (i) $\mathbf{P}(\mathbf{X} \in \Theta) = 1$.
- (ii) The set of all outcomes ω satisfying $\mathbf{X}(\omega) = x$ is an event for all $x \in \Theta$.

In plain English a discrete random variable \mathbf{X} always takes values from a countable (i.e., finite or countably infinite) subset Θ of the real numbers \mathbb{R} and the probability of \mathbf{X} being outside Θ is zero.

$$\mathbf{P}(\mathbf{X} \notin \Theta) = 0.$$

Furthermore, $\{\mathbf{X} = x\}$ is an event and hence its probability is determined by the probability law for all $x \in \Theta$. For a discrete random variable \mathbf{X} knowing $\mathbf{P}(\{\mathbf{X} = x\})$ for all $x \in \Theta$ is sufficient to calculate probabilities of all of the events that are solely about the random variable \mathbf{X} as demonstrated by the following calculation

Let S be an arbitrary subset of \mathbb{R} then

$$\begin{aligned}\mathbf{P}(\mathbf{X} \in S) &= \mathbf{P}(\mathbf{X} \in S \cap \Theta) + \mathbf{P}(\mathbf{X} \in S \setminus \Theta) && \text{by the additivity of the probability} \\ &&& \text{because } \{\mathbf{X} \in S \cap \Theta\} \cap \{\mathbf{X} \in S \setminus \Theta\} = \emptyset, \\ &= \mathbf{P}(\mathbf{X} \in S \cap \Theta) && \text{by the identity } \mathbf{P}(A) \leq \mathbf{P}(B) \text{ for } A \subset B \\ &&& \text{because } \mathbf{P}(\mathbf{X} \notin \Theta) = 0, \\ &= \sum_{x \in S \cap \Theta} \mathbf{P}(\mathbf{X} = x) && \text{by the countable additivity of the probability} \\ &&& \text{because } \{\mathbf{X} = x\} \cap \{\mathbf{X} = \tilde{x}\} = \emptyset \text{ for } x \neq \tilde{x}. \quad (6.1)\end{aligned}$$

If I am only interested in events about the random variable \mathbf{X} , i.e. events of the form $\{\mathbf{X} \in S\}$ for subsets of the real numbers S , then knowing probabilities of events $\{\mathbf{X} = x\}$ for all $x \in \Theta$ is as informative as the probability law itself. This observation motivates the following definition.

Definition 6.2. Let \mathbf{X} be a discrete random variable then the probability mass function (pmf) of \mathbf{X} is defined as

$$p_{\mathbf{X}}(x) := \mathbf{P}(\mathbf{X} = x). \quad (6.2)$$

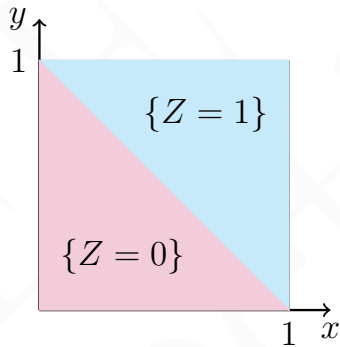
Note that

- $p_X(x) \geq 0$ by the non-negativity of the probability of any event.
- $\sum_x p_X(x) = 1$ with the tacit understanding that sum is over x values with positive $p_X(x)$ by the countable additivity of the probability.
- $\mathbf{P}(X \in S) = \sum_{x \in S} p_X(x)$ for any $S \subset \mathbb{R}$ with the tacit understanding that sum is over x values in S with positive $p_X(x)$ by the countable additivity of the probability and (6.1).

How do we calculate p_X ?

- Determine possible values of X .
- For each possible value x determine the outcomes ω satisfying $X(\omega) = x$.
- Calculate the probability of those outcomes.

Example 6.2. Let $\Omega = \{(x, y) : x \in (0, 1), y \in (0, 1)\}$ and the probability law be the continuous uniform one considered in Example 2.6.



Let the random variable Z be $Z := \lfloor x + y \rfloor$.

- Z can only take value 0 and 1.
- $p_Z(0) = \mathbf{P}(\{x + y < 1\}) = 1/2$.
- $p_Z(1) = \mathbf{P}(\{x + y \geq 1\}) = 1/2$.

Example 6.3. Let us consider two independent rolls of a fair tetrahedral die, i.e., a die in which all four faces have equal probability at each roll. Determine the probability mass functions of the random variable X, M , and S defined below.

$X(x, y) = x$

$M(x, y) = x \vee y$

$S(x, y) = x + y$

	x=1	x=2	x=3	x=4
y=1	green	red	blue	orange
y=2	green	red	blue	orange
y=3	green	red	blue	orange
y=4	green	red	blue	orange

	x=1	x=2	x=3	x=4
y=1	green	red	blue	orange
y=2	red	red	blue	orange
y=3	blue	blue	blue	orange
y=4	orange	orange	orange	orange

	x=1	x=2	x=3	x=4
y=1	green	red	blue	orange
y=2	red	blue	orange	light blue
y=3	blue	orange	light blue	yellow
y=4	orange	light blue	yellow	purple

$$p_X(x) = \begin{cases} \frac{1}{4} & \text{if } x = 1 \\ \frac{1}{4} & \text{if } x = 2 \\ \frac{1}{4} & \text{if } x = 3 \\ \frac{1}{4} & \text{if } x = 4 \\ 0 & \text{else} \end{cases}$$

$$= \begin{cases} \frac{1}{4} & \text{if } x \in \{1, \dots, 4\} \\ 0 & \text{else} \end{cases}$$

$$p_M(m) = \begin{cases} \frac{1}{16} & \text{if } m = 1 \\ \frac{3}{16} & \text{if } m = 2 \\ \frac{5}{16} & \text{if } m = 3 \\ \frac{7}{16} & \text{if } m = 4 \\ 0 & \text{else} \end{cases}$$

$$= \begin{cases} \frac{1}{4} & \text{if } z \in \{1, \dots, 4\} \\ 0 & \text{else} \end{cases}$$

$$p_S(s) = \begin{cases} \frac{1}{16} & \text{if } s = 2 \\ \frac{2}{16} & \text{if } s = 3 \\ \frac{3}{16} & \text{if } s = 4 \\ \frac{4}{16} & \text{if } s = 5 \\ \frac{3}{16} & \text{if } s = 6 \\ \frac{2}{16} & \text{if } s = 7 \\ \frac{1}{16} & \text{if } s = 8 \\ 0 & \text{else} \end{cases}$$

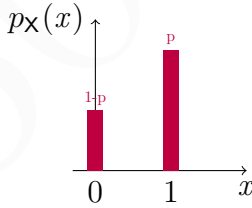
HW. Calculate the pmf of the random variables **A** and **B** defined below

$$A(x, y) = x \cdot y$$

$$B(x, y) = \frac{x}{y}$$

Certain families of probability mass functions are encountered frequently on different contexts and corresponding discrete random variables are named after the pmf. In the following, we give a partial list of such named discrete random variables.

Example 6.4 (Bernoulli Random Variable with success probability p). For $p \in [0, 1]$

$$p_X(x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \\ 0 & \text{else} \end{cases}$$


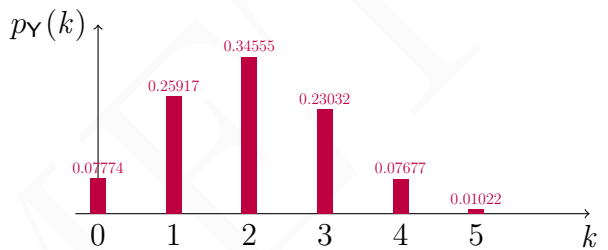
We use Bernoulli random variables to describe case where there are only two possible case: a phone line that is either busy or free, a person who is either for or against a political candidate, a person either with a disease or free of it.

Consider a random experiment whose outcome reveals the result of a coin toss among other things, i.e. the coin toss resulting in a Head and its complement, i.e., coin toss resulting in Tail, are both events. Let the random variable X be

$$X = \begin{cases} 1 & \text{if the coin toss results in a Head} \\ 0 & \text{if the coin toss results in a Tail} \end{cases}$$

If $P(\{\text{The coin toss resulting in a Head}\}) = p$ then X is a Bernoulli random variable with success probability p .

Example 6.5 (Binomial Random Variable with Parameters n and p). For $p \in [0, 1]$ and $n \in \mathbb{Z}_+$

$$p_Y(k) = \begin{cases} \binom{n}{k} p^k (1 - p)^{n-k} & \text{if } k = \{0, 1, \dots, n\} \\ 0 & \text{else} \end{cases}$$


$n = 5$ and $p = 0.4$

Consider n independent coin tosses and let the random variable Y be the total number of Heads observed in n trials.

$$\Omega = \{\omega = (\omega_1, \dots, \omega_n) : \omega_j \in \{H, T\}\}$$

$$Y(\omega) = \text{The total \# of } H\text{'s in the string } \omega \quad \forall \omega \in \Omega, \tag{6.3}$$

$$Y = \text{The total \# of heads in } n \text{ coin tosses.} \tag{6.4}$$

Note that pointwise description in (6.3) and the one in (6.4) are the same. If for each coin toss probability of getting a Head is p then probability of observing k heads has already been calculated in Section 4.4 to be the binomial probability $\binom{n}{k} p^k (1 - p)^{n-k}$ thus Y is a Binomial random variable with parameters n and p .

In the above experiment, considering events of getting a Head in each coin toss, we can define n different Bernoulli random variables:

$$\begin{aligned}
 X_j(\omega) &= \begin{cases} 1 & \text{if the } j^{\text{th}} \text{ entry of } \omega, \text{ i.e. } \omega_j, \text{ is an } H \\ 0 & \text{if the } j^{\text{th}} \text{ entry of } \omega, \text{ i.e. } \omega_j, \text{ is an } T \end{cases} & \forall \omega \in \Omega, \quad \forall j \in \{1, \dots, n\}, \\
 X_j &= \begin{cases} 1 & \text{if the } j^{\text{th}} \text{ coin toss results in a Head} \\ 0 & \text{if the } j^{\text{th}} \text{ coin toss results in a Tail} \end{cases} & \forall j \in \{1, \dots, n\}.
 \end{aligned}$$

Note that X_j 's are Bernoulli random variables with success probability p . Furthermore, as an immediate consequence of the construction, we can express $Y(\omega)$ in terms of $X_j(\omega)$'s as follows:

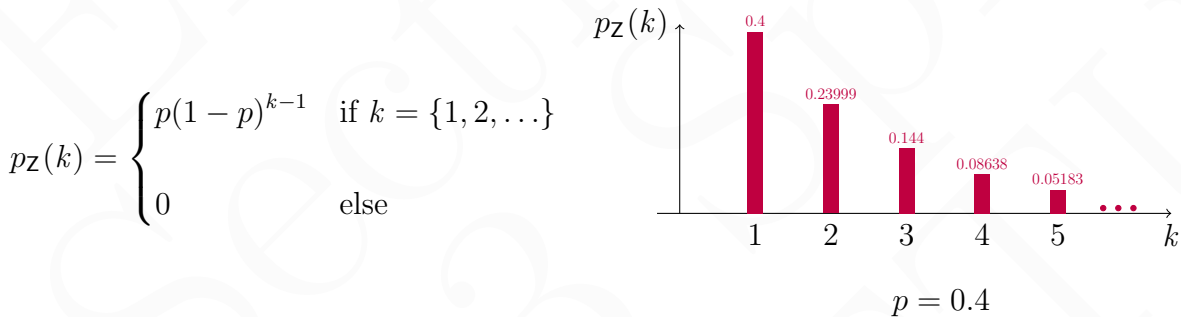
$$Y(\omega) = \sum_{j=1}^n X_j(\omega) \quad \forall \omega \in \Omega. \tag{6.5}$$

Thus the values of the functions on the left and the right of the equality sign are equal to each other for all outcomes, i.e. for all possible arguments of these function. As it is the case for functions, for random variables this is usually expressed simply as follows:

$$Y = \sum_{j=1}^n X_j. \tag{6.6}$$

Note that the equality claimed in (6.6) is not the equality of two real numbers but two functions on the sample space, to be precise the equality of one function to a sum of n others.

Example 6.6 (Geometric Random Variable with success probability p). For $p \in (0, 1)$



Consider a countably infinite sequence of coin tosses such that any finite collection of them form independent Bernoulli trials. Using the multiplication rule discussed in Section 3.3 we can describe the probability law for such a probability space, by starting from the case of n independent coin tosses adding independent coin tosses to the model one by-one-by. The subtlety here is that since we have countably infinite stages we will never be able to finish the description of the model, however, we can calculate the probabilities of any event that can be described by⁹ events about finite collections of these Bernoulli trials.

Note that possible outcomes of this random experiment are strings of letters H and T indexed by the positive integers.

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

$$Z(\omega) = \begin{cases} k & \text{if the index of the first } H \text{ in the string } \omega \text{ is } k \\ 0 & \text{if the string } \omega \text{ does not have any } H \text{ in it} \end{cases} \quad \forall \omega \in \Omega, \tag{6.7}$$

$$Z = \begin{cases} k & \text{the first } k - 1 \text{ tosses results in a T and the } k^{\text{th}} \text{ toss is an H} \\ 0 & \text{if } \omega \notin \bigcup_{k \in \mathbb{Z}_+} A_k, \text{ i.e., all of the coin toss results in a T} \end{cases}. \tag{6.8}$$

⁹Described via countably many unions, intersections, and complementations.

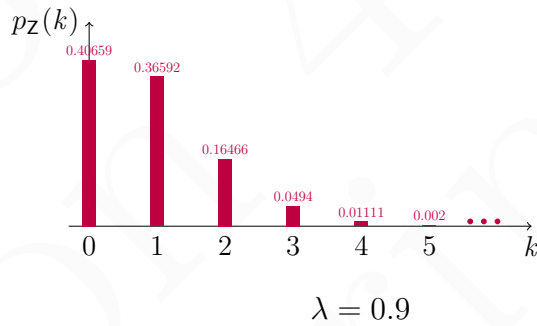
The probability of the event $\{Z = k\}$ can be calculated by considering first k coin tosses to be $\mathbf{P}(Z = k) = p(1 - p)^{k-1}$ for any $k \in \mathbb{Z}_+$. Furthermore, the probability of the event $\{Z = 0\}$ is zero, i.e., $\mathbf{P}(Z = 0) = 0$. Thus Z described (6.7) is a Geometric random variable with success probability p .

Remark 6.1. Note that $\mathbf{P}(\{\omega\}) = 0$ for all $\omega \in \Omega$ and for each positive integer k the event $\{Z(\omega) = k\}$ has uncountably many ω 's.

HW. Consider two dimensional continuous uniform probability law discussed in Example 2.6 with the outcomes of the form $\omega = (x, y)$ and let the random variable S be the index of the first 1 in the binary representation of x . Is S a geometric random variable?

Example 6.7 (Poisson Random Variable with parameter λ). For $\lambda \in \mathbb{R}_+$

$$p_X(k) = \begin{cases} e^{-\lambda} \frac{\lambda^k}{k!} & \text{if } k = \{0, 1, 2, \dots\} \\ 0 & \text{else} \end{cases}$$

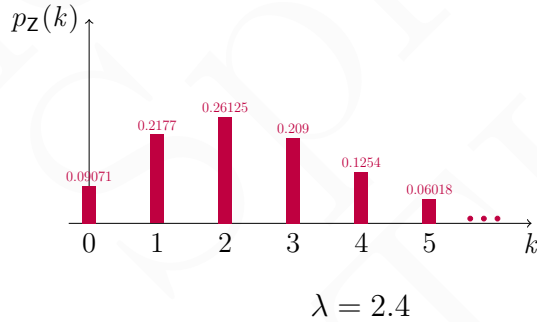


The Taylor expansion of the exponential function around 0 is

$$e^\lambda = 1 + \frac{\lambda}{1!} + \frac{\lambda^2}{2!} + \dots = \sum_{k \in \mathbb{Z}_{\geq 0}} \frac{\lambda^k}{k!}$$

Thus

$$\sum_{k \in \mathbb{Z}_{\geq 0}} p_X(k) = e^{-\lambda} e^\lambda = 1.$$



As it should be.

Poisson probability mass function provide a good approximation to Binomial probability mass function when $n \gg k$ and $p = \frac{\lambda}{n}$, i.e.

$$\binom{n}{k} p^k (1 - p)^{n-k} \approx e^{-\lambda} \frac{\lambda^k}{k!} \quad \text{for } p = \frac{\lambda}{n} \text{ and } n \gg k$$

Why?

$$\begin{aligned} \binom{n}{k} p^k (1 - p)^{n-k} &= \frac{1}{k!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \prod_{\ell=0}^{k-1} (n - \ell) \\ &= \frac{\lambda^k}{k!} \left(1 - \frac{\lambda}{n}\right)^{n-k} \prod_{\ell=0}^{k-1} \frac{n - \ell}{n} \end{aligned}$$

On the other hand for fixed k and λ that does not change with n we have

$$\begin{aligned} \lim_{n \rightarrow \infty} \prod_{\ell=0}^{k-1} \frac{n - \ell}{n} &= 1, \\ \lim_{n \rightarrow \infty} \left(1 - \frac{\lambda}{n}\right)^{n-k} &= e^{-\lambda}. \end{aligned}$$

The first limit is easy to see, for the second limit one can use l'Hospital's rule after taking the natural logarithm of the expression.

6.2 Functions of A Random Variable

Let $X(\cdot)$ be a discrete random variable taking values in a countable set Θ and $g(\cdot)$ be a function that is defined on Θ , then $g(X(\cdot))$ is a discrete random variable. In other words $g \circ X : \Omega \rightarrow \mathbb{R}$ is a discrete random variable for any function $g : \Theta \rightarrow \mathbb{R}$ and discrete random variable $X : \Omega \rightarrow \Theta$. To establish this assertion we need to find a countable set Φ satisfying the following two conditions

- $P(g(X) \in \Phi) = 1$
- $\{\omega \in \Omega : g(X(\omega)) = y\}$ is an event for all $y \in \Phi$.

The image of Θ under g , i.e. $g(\Theta) = \{y = g(x) : x \in \Theta\}$, is a countable set. Furthermore, $\{\omega \in \Omega : g(X(\omega)) = y\} = \bigcup_{x:g(x)=y} \{\omega \in \Omega : X(\omega) = x\}$ is an event for all $y \in g(\Theta)$ because countable unions of events are events. Thus we can choose Φ to be $g(\Theta)$. Then

$$Y = g(X) \quad \Rightarrow \quad p_Y(y) = \sum_{x:g(x)=y} p_X(x) \quad (6.9)$$

Example 6.8. Let us consider two independent rolls of a fair tetrahedral die, i.e., a die in which all four faces have equal probability at each roll. Let the random variable D be

$$D(x, y) = x - y$$

Determine the probability mass functions of random variables $|D|$ and D^2 .

$D(x, y) = x - y$

	x=1	x=2	x=3	x=4
y=1				
y=2				
y=3				
y=4				

$$p_D(k) = \begin{cases} \frac{4-|k|}{16} & \text{if } k \in \{-3, -2, -1, 0, 1, 2, 3\} \\ 0 & \text{else} \end{cases}$$

$|D|(x, y) = |x - y|$

	x=1	x=2	x=3	x=4
y=1				
y=2				
y=3				
y=4				

$$p_{|D|}(k) = \begin{cases} \frac{3-|k-1|}{8} & \text{if } k \in \{0, 1, 2, 3\} \\ 0 & \text{else} \end{cases}$$

$D^2(x, y) = (x - y)^2$

	x=1	x=2	x=3	x=4
y=1				
y=2				
y=3				
y=4				

$$p_{D^2}(k) = \begin{cases} \frac{3-|\sqrt{k}-1|}{8} & \text{if } k \in \{0, 1, 4, 9\} \\ 0 & \text{else} \end{cases}$$

6.3 Expected Value and Variance of a Random Variable

Definition 6.3 (Expected Value/Expectation/Mean). The expected value of a discrete random variable X is defined as

$$\mathbf{E}[X] := \sum_x x p_X(x) \tag{6.10}$$

When the PMF of X is non-zero only for finitely many x values there are only finitely many terms in the sum (6.10) and its meaning is clear. If there are infinitely many terms in the sum, we can calculate the contribution of positive and negative terms separately,

$$\mathbf{E}[X^+] := \sum_{x:x>0} x p_X(x), \quad \mathbf{E}[X^-] := \sum_{x:x<0} (-x) p_X(x), \tag{6.11}$$

The sum defining $\mathbf{E}[X^+]$ is composed of positive terms only, thus either the sum converges to a real number or it diverges to infinity; and this conclusion is irrespective of the order we add the terms in this infinite sum. The same assertions are true for $\mathbf{E}[X^-]$. Then

$$\mathbf{E}[X] = \begin{cases} \mathbf{E}[X^+] - \mathbf{E}[X^-] & \text{if either } \mathbf{E}[X^+] < \infty \text{ or } \mathbf{E}[X^-] < \infty, \\ \text{undefined} & \text{if } \mathbf{E}[X^+] = \infty \text{ and } \mathbf{E}[X^-] = \infty. \end{cases}$$

For the cases when expectation $\mathbf{E}[X]$ is defined the order in which we add the terms (6.10) does not matter we get the same value. This, however, is not true when both $\mathbf{E}[X^+]$ and $\mathbf{E}[X^-]$ are infinite.

Example 6.9. Let the pmf of the discrete random variable X be

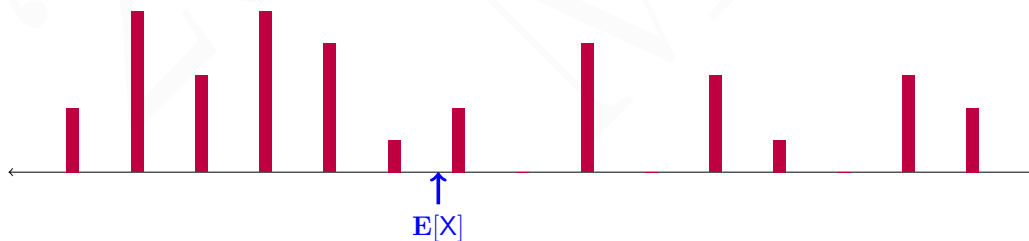
$$p_X(x) = \begin{cases} 3^{-k} & \text{if } x = \{-3^k, 3^k\} \text{ and } k \in \mathbb{Z}^+ \\ 0 & \text{else} \end{cases}$$

Calculate $\mathbf{E}[X^+]$ and $\mathbf{E}[X^-]$.

$$\begin{aligned} \sum_x p_X(x) &= 2 \sum_{k=1}^{\infty} 3^{-k} &= \frac{2}{3} \sum_{k=0}^{\infty} 3^{-k} &= \frac{2}{3} \frac{1}{1 - \frac{1}{3}} &= 1. \\ \mathbf{E}[X^+] &= \sum_{k=1}^{\infty} 3^{-k} 3^k &= \sum_{k=1}^{\infty} 1 &= \infty \\ \mathbf{E}[X^-] &= \sum_{k=1}^{\infty} 3^{-k} 3^k &= \sum_{k=1}^{\infty} 1 &= \infty \end{aligned}$$

Note that in the partial sums of (6.10) can be made arbitrarily large with positive sign by taking more positive terms and arbitrarily large with negative sign by taking more negative terms. A similar situation occurs whenever both $\mathbf{E}[X^+]$ and $\mathbf{E}[X^-]$ are infinite.

In the following we will focus exclusively to the case when X has a finite expected value, i.e. to the case when both $\mathbf{E}[X^+]$ and $\mathbf{E}[X^-]$ are finite.



The expected value can be interpreted as the center of mass (the balance point, the center of gravity), point with respect to which torque is zero.

The expected value is also what we “expect” the *empirical mean* (sample mean) of many independent trials of a random variable will behave like.

Let the amount of reward a player get from a game be a random variable X with the possible values of x_1, x_2, \dots, x_k . Let $\#_N(x_j)$ be the number of times the player gets the reward x_j in N games, which are assumed to be independent trials. Then we expect

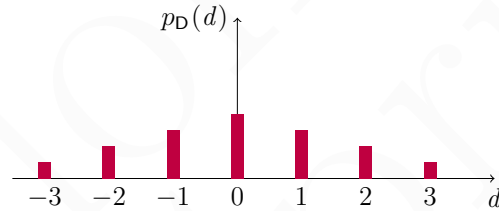
$$\underbrace{\frac{x_1\#_N(x_1) + x_2\#_N(x_2) + \dots + x_k\#_N(x_k)}{N}}_{\text{empirical mean}} \approx \underbrace{x_1p_X(x_1) + x_2p_X(x_2) + \dots + x_kp_X(x_k)}_{\text{expected value}}$$

The empirical mean is a random variable and the expected value is a real number. What we mean by \approx is that for large N with very high probability the empirical mean will be in the vicinity of the expected values. We will make this claim precise while discussing the law of large numbers near the end of the semester.

Example 6.10. Let us calculate the expected value of D , $|D|$, and D^2 discussed in Example 6.8

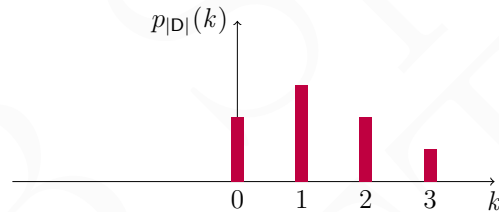
$$p_D(d) = \begin{cases} \frac{4-|d|}{16} & \text{if } d \in \mathbb{Z} \ \& \ |d| \leq 3 \\ 0 & \text{else} \end{cases}$$

$$\mathbf{E}[D] = \sum_{d=-3}^3 d \frac{4-|d|}{16} = 0$$



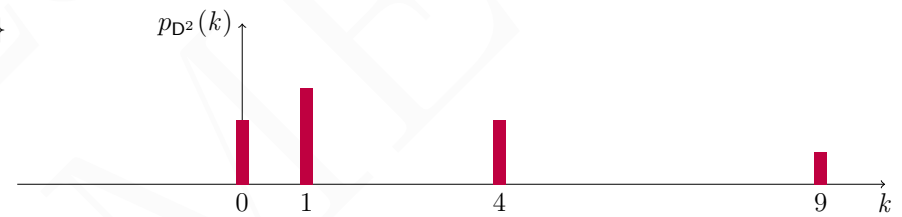
$$p_{|D|}(k) = \begin{cases} \frac{3-|k-1|}{8} & \text{if } k \in \{0, 1, 2, 3\} \\ 0 & \text{else} \end{cases}$$

$$\mathbf{E}[|D|] = \sum_{k=0}^3 k \frac{3-|k-1|}{8} = 1.25$$



$$p_{D^2}(k) = \begin{cases} \frac{3-|\sqrt{k}-1|}{8} & \text{if } k \in \{0, 1, 4, 9\} \\ 0 & \text{else} \end{cases}$$

$$\mathbf{E}[D^2] = \sum_{k \in \{0,1,4,9\}} k \frac{3-|\sqrt{k}-1|}{8} = 2.5$$



Note that

$$\begin{aligned} \mathbf{E}[|D|] &= \sum_{d=-3}^3 |d|p_D(d) &= \sum_{d=-3}^3 |d| \frac{4-|d|}{16} &= 1.25 \\ \mathbf{E}[D^2] &= \sum_{d=-3}^3 d^2p_D(d) &= \sum_{d=-3}^3 d^2 \frac{4-|d|}{16} &= 2.5 \\ \mathbf{E}[D^2] &= \sum_{\ell=0}^3 \ell^2 p_{|D|}(\ell) &= \sum_{\ell=0}^3 \ell^2 \frac{3-|\ell-1|}{8} &= 2.5 \end{aligned}$$

Expected Value of a Function of Random Variable. If a random variable discrete Y satisfies $Y = g(X)$ for another discrete random variable X and function $g(\cdot)$, then

$$\mathbf{E}[Y] = \sum_x g(x)p_X(x). \quad (6.12)$$

$$= \mathbf{E}[g(X)] \quad (6.13)$$

Proof.

$$\begin{aligned} \mathbf{E}[Y] &= \sum_y y p_Y(y) \\ &= \sum_y y \sum_{x:g(x)=y} p_X(x) && \text{by } p_Y(y) = \sum_{x:g(x)=y} p_X(x) \\ &= \sum_y \sum_{x:g(x)=y} g(x)p_X(x) \\ &= \sum_x g(x)p_X(x) \end{aligned}$$

□

Note that (6.12) asserts that we can use the pmf of the random variable X to calculate the expected value of the random variable $Y = g(X)$. (6.12) DOES NOT assert that expected value of $g(X)$ is the value of g at $\mathbf{E}[X]$. Thus ~~$\mathbf{E}[g(X)] = g(\mathbf{E}[X])$~~ as demonstrated by the following example.

Example 6.11. A student's daily commute is 5km. Each day she decides whether to walk at a pace of 6km per hour or ride her bike at a pace of 15km per hour. The probability of her walking for the day is $1/3$, and the probability of her using her bike for the day is $2/3$.

- (a) What is the expected pace of her commute?
 (b) What is the expected commuting time for her?
 (a) Let V be the random variable describing the pace in km per hour. Then

$$\begin{aligned} p_V(v) &= \begin{cases} \frac{1}{3} & v = 6 \\ \frac{2}{3} & v = 15 \end{cases} \\ \mathbf{E}[V] &= \frac{1}{3} \cdot 6 + \frac{2}{3} \cdot 15 = 12 \end{aligned}$$

- (b) Let T be the random variable describing commuting time for her.

$$\begin{aligned} T &= \frac{5}{V} \\ \mathbf{E}[T] &= \frac{1}{3} \cdot \frac{5}{6} + \frac{2}{3} \cdot \frac{5}{15} = \frac{1}{2} \end{aligned}$$

Note that $T = \frac{5}{V}$ but $\mathbf{E}[T] \neq \frac{5}{\mathbf{E}[V]}$.

Definition 6.4.

$\mathbf{E}[X^n]$ is called the n^{th} moment of the r.v. X .

$\text{var}(X) := \mathbf{E}[(X - \mathbf{E}[X])^2]$ is called the variance of the r.v. X .

$\sigma_X := \sqrt{\text{var}(X)}$ is called the standard deviation of the r.v. X .

Variance and Second Moments of a Random Variable.

$$\text{var}(X) = \mathbf{E}[X^2] - (\mathbf{E}[X])^2 \quad (6.14)$$

$$\mathbf{E}[(X - c)^2] = \text{var}(X) + (\mathbf{E}[X] - c)^2 \quad \forall c \in \mathbb{R}. \quad (6.15)$$

Proof.

$$\begin{aligned} \mathbf{E}[(X - c)^2] &= \sum_x (x - c)^2 p_X(x) \\ &= \sum_x (x^2 - 2xc + c^2) p_X(x) \\ &= \sum_x x^2 p_X(x) - 2c \cdot \sum_x x p_X(x) + c^2 \cdot \sum_x p_X(x) \\ &= \mathbf{E}[X^2] - 2c \cdot \mathbf{E}[X] + c^2 \cdot 1. \end{aligned} \quad (6.16)$$

(6.14) follows from (6.16) and the definition of variance by setting $c = \mathbf{E}[X]$.(6.15) follows from (6.14) and (6.16) □

Remark 6.2. An function g is said to be linear if it is of the form $g(x) = a \cdot x + b$ for some constants a and b . In certain contexts only $b = 0$ case is called linear and non-zero b case is called affine. We will not make that distinction in in EE230, non-zero b cases will be called linear as well.

Linear Function of a Random Variable. If $Y = aX + b$ for some constants a and b then

$$\mathbf{E}[Y] = a\mathbf{E}[X] + b \quad (6.17)$$

$$\text{var}(Y) = a^2 \text{var}(X) \quad (6.18)$$

Proof.

$$\begin{aligned} \mathbf{E}[Y] &= \mathbf{E}[aX + b] \\ &= \sum_x (ax + b) \cdot p_X(x). && \text{by (6.12)} \\ &= a \cdot \sum_x x p_X(x) + b \cdot \sum_x p_X(x) \\ &= a\mathbf{E}[X] + b. \\ \text{var}(Y) &= \mathbf{E}[(Y - \mathbf{E}[Y])^2] \\ &= \mathbf{E}[(aX + b - (a\mathbf{E}[X] + b))^2] && \text{by (6.17)} \\ &= \mathbf{E}[a^2(X - \mathbf{E}[X])^2] \\ &= a^2 \mathbf{E}[(X - \mathbf{E}[X])^2] && \text{by applying (6.17) for } (X - \mathbf{E}[X])^2 \\ &= a^2 \text{var}(X) \end{aligned}$$

□

Example 6.12 (Bernoulli Random Variable with success probability p). For $p \in [0, 1]$

$$\begin{aligned} p_X(x) &= \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases} && \Rightarrow && \mathbf{E}[X] = p \\ X^2 &= X && \Rightarrow && \mathbf{E}[X^2] = p \\ \text{var}(X) &= \mathbf{E}[X^2] - (\mathbf{E}[X])^2 && \Rightarrow && \text{var}(X) = p(1 - p) = p - p^2 \end{aligned}$$

Example 6.13 (Discrete Uniform Random Variable).

$$p_X(k) = \begin{cases} \frac{1}{n+1} & k \in \{0, 1, \dots, n\} \\ 0 & \text{else} \end{cases}$$

$$\begin{aligned} \mathbf{E}[X] &= \frac{1}{n+1} \sum_{k=0}^n k \\ &= \frac{1}{n+1} \frac{(n+1)n}{2} \\ &= \frac{n}{2} \end{aligned}$$

$$\begin{aligned} \mathbf{E}[X^2] &= \frac{1}{n+1} \sum_{k=0}^n k^2 \\ &= \frac{1}{n+1} \frac{n \cdot (n+1) \cdot (2n+1)}{6} \quad \text{by } 1 + 2^2 + \dots + n^2 = \frac{n(n+1)(2n+1)}{6} \\ &= \frac{n \cdot (2n+1)}{6} \end{aligned}$$

$$\begin{aligned} \text{var}(X) &= \mathbf{E}[X^2] - (\mathbf{E}[X])^2 \\ &= \frac{n \cdot (2n+1)}{6} - \frac{n^2}{4} \\ &= \frac{n \cdot (n+2)}{12} \end{aligned}$$

HW. Prove $1 + \dots + n^2 = \frac{n \cdot (n+1) \cdot (2n+1)}{6}$ by induction: evidently identity holds for $n = 1$. Assume $\sum_{k=1}^n k^2 = \frac{n \cdot (n+1) \cdot (2n+1)}{6}$ prove $\sum_{k=1}^{n+1} k^2 = \frac{(n+1) \cdot ((n+1)+1) \cdot (2 \cdot (n+1)+1)}{6}$.

HW. Determine the expected value and the variance of the random variable Z with the following probability mass function.

$$p_Z(k) = \begin{cases} \frac{1}{b-a+1} & k \in \{a, a+1, \dots, b\} \\ 0 & \text{else} \end{cases}$$

Example 6.14 (Poisson Random Variable with parameter λ). For $\lambda \in \mathbb{R}_+$

$$p_X(k) = \begin{cases} e^{-\lambda} \frac{\lambda^k}{k!} & k \in \{0, 1, 2, \dots\} \\ 0 & \text{else} \end{cases}$$

$$\begin{aligned} \mathbf{E}[X] &= \sum_{k=0}^{\infty} k e^{-\lambda} \frac{\lambda^k}{k!} \\ &= e^{-\lambda} \sum_{k=1}^{\infty} k \frac{\lambda^k}{k!} \\ &= e^{-\lambda} \cdot \lambda \cdot \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \\ &= e^{-\lambda} \cdot \lambda \cdot e^{\lambda} \\ &= \lambda \end{aligned}$$

$$\begin{aligned} \text{var}(X) &= \mathbf{E}[X^2] - (\mathbf{E}[X])^2 \\ &= \lambda \end{aligned}$$

$$\text{by } e^{\lambda} = \sum_{k \in \mathbb{Z}_{\geq 0}} \frac{\lambda^k}{k!}$$

$$\begin{aligned} \mathbf{E}[X^2] &= \sum_{k=0}^{\infty} k^2 e^{-\lambda} \frac{\lambda^k}{k!} \\ &= e^{-\lambda} \sum_{k=1}^{\infty} k \frac{\lambda^k}{(k-1)!} \\ &= e^{-\lambda} \lambda \sum_{\ell=0}^{\infty} (\ell+1) \frac{\lambda^{\ell}}{\ell!} \\ &= e^{-\lambda} \lambda \left(\sum_{\ell=0}^{\infty} \ell \frac{\lambda^{\ell}}{\ell!} + \sum_{\ell=0}^{\infty} \frac{\lambda^{\ell}}{\ell!} \right) \\ &= e^{-\lambda} \lambda \left(\lambda \sum_{\ell=1}^{\infty} \frac{\lambda^{\ell-1}}{(\ell-1)!} + \sum_{\ell=0}^{\infty} \frac{\lambda^{\ell}}{\ell!} \right) \\ &= \lambda(1 + \lambda) \end{aligned}$$

6.4 Joint Probability Mass Function

Random variables allow us to analyze probabilistic models (i.e., probability spaces) without explicitly referring to the underlying sample space and the probability law. When reasoning about a single discrete random variable X , its probability mass function p_X is enough, as we have demonstrated in (6.1). The probability mass function p_X tells everything there is to know about X , but in isolation. When reasoning about two random variables X and Y , knowing the probability mass functions p_X and p_Y is not enough, we need to know their joint behavior. The joint probability mass functions we discuss in the following serves this purpose.

For expressing probabilities of intersections of events about random variables, we use the following shorthand representation

$$\begin{aligned} \mathbf{P}\left(\{\omega \in \Omega : X(\omega) = x\} \cap \{\omega \in \Omega : Y(\omega) = y\}\right) &= \mathbf{P}(\{\omega \in \Omega : X(\omega) = x \text{ and } Y(\omega) = y\}), \\ &= \mathbf{P}(\{X = x \text{ and } Y = y\}), \\ &= \mathbf{P}\left(\{X = x\} \cap \{Y = y\}\right), \\ &= \mathbf{P}(\{X = x, Y = y\}), \\ &= \mathbf{P}(X = x, Y = y). \end{aligned}$$

Definition 6.5. Let X_1, X_2, \dots, X_n be discrete random variables then the joint probability mass function of X_1, X_2, \dots, X_n is

$$p_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) := \mathbf{P}(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n). \quad (6.19)$$

Then for two random variables X and Y the joint pmf of X and Y is given by

$$p_{X,Y}(x, y) = \mathbf{P}(X = x, Y = y). \quad (6.20)$$

We can calculate the probability of any event about the random variables X and Y using their joint pmf, i.e.,

$$\mathbf{P}((X, Y) \in S) = \sum_{(x,y) \in S} p_{X,Y}(x, y). \quad (6.21)$$

This is in a sense nothing more than a restatement of (6.1), taking into account the fact that the Cartesian product of two countable sets is countable. In particular, by considering events of the form $S_a = \{(x, y) : x = a\}$ for different a values we can determine the pmf of X , i.e.

$$\begin{aligned} p_X(a) &= \mathbf{P}(X = a) && \text{by the definition of } p_X \\ &= \mathbf{P}((X, Y) \in S_a) && \text{by the definition of } S_a \\ &= \sum_{(x,y) \in S_a} p_{X,Y}(x, y) && \text{by (6.21)} \\ &= \sum_{(x,y): x=a} p_{X,Y}(x, y) && \text{by the definition of } S_a \\ &= \sum_y p_{X,Y}(a, y). \end{aligned}$$

Similarly, pmf of Y is also determined by the joint pmf of X and Y :

$$p_Y(y) = \sum_x p_{X,Y}(x, y).$$

p_X and p_Y are called the marginal pmfs. More generally the joint pmf of X_1, X_2, \dots, X_n , determines the joint pmf any group of random variable from X_1, X_2, \dots, X_n , e.g.,

$$\begin{aligned} p_{X_3, \dots, X_{n-1}}(x_3, \dots, x_{n-1}) &= \sum_{x_1, x_2, x_n} p_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n), \\ p_{X_k}(x_k) &= \sum_{x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n} p_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n). \end{aligned}$$

Example 6.15. Let us consider two independent rolls of a fair tetrahedral die, i.e., a die in which all four faces have equal probability at each roll. Let the random variables T and M be

$$T(x, y) = |x - y| \qquad M(x, y) = x \vee y$$

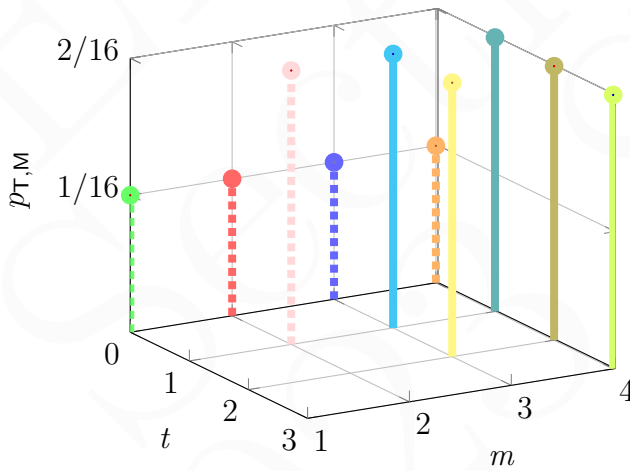
- (a) Determine the joint probability mass function of the random variables T and M , i.e., $p_{T,M}$.
- (b) Determine T and M marginals, i.e., p_T and p_M .
- (c) Determine $\mathbf{P}(T \cdot M \geq 7)$. Is it possible to determine $\mathbf{P}(T \cdot M \geq 7)$ relying solely on the marginals?

(a) $T(x, y) = x - y $	x=1	x=2	x=3	x=4
y=1				
y=2				
y=3				
y=4				

$M(x, y) = x \vee y$	x=1	x=2	x=3	x=4
y=1				
y=2				
y=3				
y=4				

$(t, m) = (x - y , x \vee y)$	x=1	x=2	x=3	x=4
y=1	(0,1)	(1,2)	(2,3)	(3,4)
y=2	(1,2)	(0,2)	(1,3)	(2,4)
y=3	(2,3)	(1,3)	(0,3)	(1,4)
y=4	(3,4)	(2,4)	(1,4)	(0,4)

$$p_{T,M}(t, m) = \begin{cases} \frac{1}{16} & \text{if } t = 0, m \in \{1, 2, 3, 4\} \\ \frac{2}{16} & \text{if } t = 1, m \in \{2, 3, 4\} \\ \frac{2}{16} & \text{if } t = 2, m \in \{3, 4\} \\ \frac{2}{16} & \text{if } t = 3, m = 4 \\ 0 & \text{else} \end{cases}$$

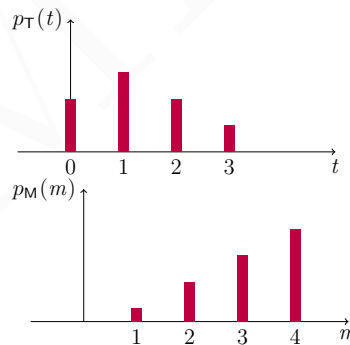


	m = 1	m = 2	m = 3	m = 4
t = 0	1/16	1/16	1/16	1/16
t = 1	0	2/16	2/16	2/16
t = 2	0	0	2/16	2/16
t = 3	0	0	0	2/16

We can apply the tabular method to calculate the marginals. The row sums will give the pmf of T ; the column sums will give the pmf of M

(b)
$$p_T(t) = \begin{cases} \frac{3-|t-1|}{8} & \text{if } t \in \{0, 1, 2, 3\} \\ 0 & \text{else} \end{cases}$$

$$p_M(m) = \begin{cases} \frac{2m-1}{16} & \text{if } m \in \{1, 2, 3, 4\} \\ 0 & \text{else} \end{cases}$$



(c)

$$\mathbf{P}(T \cdot M \geq 7) = \sum_{(t,m): t \cdot m \geq 7} p_{T,M}(t, m) = p_{T,M}(2, 4) + p_{T,M}(3, 4) = \frac{1}{4}.$$

6.5 Functions of Multiple Random Variables

Let X_1, \dots, X_n be discrete random variables then $g(X_1, \dots, X_n)$ is a discrete random variable for any function $g : \mathbb{R}^n \rightarrow \mathbb{R}$. This can be confirmed by an argument similar to the one used for establishing the same fact for $n = 1$ case in §6.2 previously. Furthermore, one can calculate the pmf of $g(X_1, \dots, X_n)$ using the joint pmf of the random variables X_1, \dots, X_n :

$$Y = g(X_1, \dots, X_n) \quad \Rightarrow \quad p_Y(y) = \sum_{(x_1, \dots, x_n): g(x_1, \dots, x_n)=y} p_{X_1, \dots, X_n}(x_1, \dots, x_n) \quad (6.22)$$

The expected value of the random variable $Y = g(X_1, \dots, X_n)$ can be calculated using its own pmf p_Y or the joint pmf of the random variables X_1, \dots, X_n :

$$\mathbf{E}[Y] = \sum_{(x_1, \dots, x_n)} g(x_1, \dots, x_n) p_{X_1, \dots, X_n}(x_1, \dots, x_n) \quad (6.23)$$

$$= \mathbf{E}[g(X_1, \dots, X_n)] \quad (6.24)$$

Note that (6.22) and (6.23) are n random variable generalizations of the identities given in (6.9) and (6.12). These n random variable generalizations follow from arguments similar to the ones used for the single random variable cases.

For linear functions (6.17) has the following generalization. For any collection of n discrete random variables X_1, \dots, X_n and $n + 1$ real numbers a_0, a_1, \dots, a_n :

$$\mathbf{E}[a_n X_n + \dots + a_1 X_1 + a_0] = a_n \mathbf{E}[X_n] + \dots + a_1 \mathbf{E}[X_1] + a_0 \quad (6.25)$$

(6.25) is often called the linearity of expectation.

HW. Confirm (6.25) using (6.23).

Remark 6.3. As you might expect (6.18) has an n random variable generalization as well, but we will it later in the semester after introducing the concept of covariance.

Example 6.16. Let the random variable Z be the difference of the random variables M and T , i.e., $Z = M - T$, and the joint pmf of the discrete random variables T and M be

$$p_{T,M}(t, m) = \begin{cases} \frac{1}{16} & \text{if } t = 0, m \in \{1, 2, 3, 4\} \\ \frac{2}{16} & \text{if } t = 1, m \in \{2, 3, 4\} \\ \frac{2}{16} & \text{if } t = 2, m \in \{3, 4\} \\ \frac{2}{16} & \text{if } t = 3, m = 4 \\ 0 & \text{else} \end{cases}$$

$t \backslash m$	1	2	3	4
0	1/16	1/16	1/16	1/16
1	0	2/16	2/16	2/16
2	0	0	2/16	2/16
3	0	0	0	2/16

(a) Calculate expected value of the random variable Z .

$$\mathbf{E}[Z] = \sum_{(t,m)} (m - t) \cdot p_{T,M}(t, m) = \frac{1}{16} (1 + 2 + 3 + 4) + \frac{2}{16} (1 + 2 + 3 + 1 + 2 + 1) = \frac{15}{8}$$

(b) Calculate the pmf of the random variable Z .

$$p_Z(z) = \begin{cases} \frac{7}{16} & \text{if } z = 1 \\ \frac{5}{16} & \text{if } z = 2 \\ \frac{3}{16} & \text{if } z = 3 \\ \frac{1}{16} & \text{if } z = 4 \\ 0 & \text{else} \end{cases}$$

Example 6.17. n students are taking a class and for each student probability of gets the grade AA with probability p . What is the expected number of students that gets the grade AA.

Let X_i be a random variable that is equal to one if i^{th} student gets the grade AA and equals to zero otherwise, i.e.

$$X_i = \begin{cases} 1 & \text{if } i^{\text{th}} \text{ student gets the grade AA} \\ 0 & \text{else} \end{cases}$$

Then the number of students who gets the grade AA is $X_1 + \dots + X_n$, i.e.,

$$X = \sum_{i=1}^n X_i$$

Then using first the linearity of the expectation and then the fact that X_i 's are Bernoulli random variables with success probability p we get

$$\begin{aligned} \mathbf{E}[X] &= \mathbf{E}\left[\sum_{i=1}^n X_i\right] \\ &= \sum_{i=1}^n \mathbf{E}[X_i] \\ &= np \end{aligned}$$

Note that we have not made any assumption about the joint pmf p_{X_1, \dots, X_n} of the random variables X_1, \dots, X_n , we because we are not given any information about it. p_{X_1, \dots, X_n} can be any joint pmf whose marginals are all Bernoulli pmfs with success probability p . Two “extreme” cases are the followings:

- All of the students are getting the same grade and thus, i.e. $X_1 = X_i$ for all $i \in \{1, \dots, n\}$. In this case X will satisfy $X = nX_1$ and its pmf is given by

$$p_X(x) = \begin{cases} p & \text{if } x = n \\ 1 - p & \text{if } x = 0 \\ 0 & \text{else} \end{cases}$$

- The students getting the grade AA, forms independent trials of a Bernoulli trial with success probability p . Then X will be a Binomial random variable with the following pmf, as discussed in Example 6.5

$$p_X(x) = \begin{cases} \binom{n}{x} p^x (1 - p)^{n-x} & \text{if } x \in 0, 1, \dots, n \\ 0 & \text{else} \end{cases}$$

Example 6.18. Attendees of a party are required to have a hat; n attendees of this party throw their hats into a bag and then pick one hat each from the bag. All assignment of n hats to n attendees are equally likely and X is the number of attendees who get their own hat back. Determine $\mathbf{E}[X]$.

$$X = \sum_{i=1}^n X_i \quad \text{where} \quad X_i = \begin{cases} 1 & \text{if } i^{\text{th}} \text{ individual gets their hat back} \\ 0 & \text{else} \end{cases}$$

Note that X_i 's are Bernoulli random variables with success probability $1/n$, then $\mathbf{E}[X] = 1$.

Remark 6.4. Given a collection of random variables X_1, \dots, X_n instead of considering one function g one can consider a collection of functions g_1, \dots, g_k and obtain the joint pmf of the corresponding random variables,

$$Y_i = g_i(X_1^n) \quad \forall i \in \{1, \dots, \kappa\} \quad \Rightarrow \quad p_{Y_1^\kappa}(y_1^\kappa) = \sum_{x_1^n: g_i(x_1^n)=y_i \quad \forall i \in \{1, \dots, \kappa\}} p_{X_1^n}(x_1^n), \quad (6.26)$$

where $x_1^n = (x_1, \dots, x_n)$, $y_1^\kappa = (y_1, \dots, y_\kappa)$, etc.

6.6 Conditioning

6.6.1 Conditional PMFs Given an Event

For a given probability space $(\Omega, \mathcal{F}, \mathbf{P}(\cdot))$ and an event $A \in \mathcal{F}$ with positive probability, i.e., $\mathbf{P}(A) > 0$, the conditional probability law $\mathbf{P}(\cdot|A)$ is defined in §3.2 as

$$\mathbf{P}(B|A) := \frac{\mathbf{P}(B \cap A)}{\mathbf{P}(A)} \quad \forall B \in \mathcal{F}$$

In §3.2, we have confirmed that $(\Omega, \mathcal{F}, \mathbf{P}(\cdot|A))$ is a probability space. Since the initial probability law $\mathbf{P}(\cdot)$ and the conditional probability law $\mathbf{P}(\cdot|A)$ share the same sample space Ω and the same σ -algebra any random variable defined on the probability space $(\Omega, \mathcal{F}, \mathbf{P}(\cdot))$ is a random variable for the conditional probability space $(\Omega, \mathcal{F}, \mathbf{P}(\cdot|A))$, as well.

For any discrete random variable X and any event A the conditional pmf $p_{X|A}$ is defined as

$$p_{X|A}(x) := \mathbf{P}(X = x|A) \quad \forall x. \tag{6.27}$$

Since X is a discrete random variable there is a countable set Θ such that $\mathbf{P}(X \notin \Theta) = 0$. Then

$$\begin{aligned} \mathbf{P}(X \notin \Theta|A) &= \frac{\mathbf{P}(\{X \notin \Theta\} \cap A)}{\mathbf{P}(A)} && \text{by the definition of conditional probability} \\ &= 0 && \text{because } \mathbf{P}(\{X \notin \Theta\} \cap A) \leq \mathbf{P}(\{X \notin \Theta\}) = 0 \\ \sum_{x \in \Theta} p_{X|A}(x) &= \mathbf{P}(X \in \Theta|A) && \text{by the } \sigma\text{-additivity for } \mathbf{P}(\cdot|A) \\ &= 1 - \mathbf{P}(X \notin \Theta|A) \\ &= 1 && \text{because } \mathbf{P}(X \notin \Theta|A) = 0. \end{aligned}$$

As one would expect. In a sense the pmf of a random variable is its probability law and the conditional pmf of the random variable is the conditional probability law. We define the expected value of the random variable X given the event A in terms of the conditional pmf as follows

$$\mathbf{E}[X|A] := \sum_x x p_{X|A}(x) \tag{6.28}$$

Example 6.19. Let us consider two independent rolls of a fair tetrahedral die. Let the random variable T be the absolute value of the difference of results of two rolls and the event A the maximum of two rolls being equal to 4:

$$T(x, y) = |x - y| \quad A = \{x \vee y = 4\}$$

Determine the conditional probability mass functions $p_{T|A}$ and $p_{T|A^c}$. Determine the conditional probability of the event A given that $T = 2$.

$T(x, y) = x - y $	x=1	x=2	x=3	x=4
y=1	A ^c	A ^c	A ^c	A
y=2	A ^c	A ^c	A ^c	A
y=3	A ^c	A ^c	A ^c	A
y=4	A	A	A	A

$$p_T(t) = \begin{cases} 1/4 & \text{if } t \in \{0, 2\} \\ 3/8 & \text{if } t = 1 \\ 1/8 & \text{if } t = 3 \\ 0 & \text{else} \end{cases}$$

$$\mathbf{P}(A) = \frac{7}{16} \quad p_{T|A}(t) = \begin{cases} 1/7 & \text{if } t = 0 \\ 2/7 & \text{if } t \in \{1, 2, 3\} \\ 0 & \text{else} \end{cases} \quad p_{T|A^c}(t) = \begin{cases} 1/3 & \text{if } t = 0 \\ 4/9 & \text{if } t = 1 \\ 2/9 & \text{if } t = 2 \\ 0 & \text{else} \end{cases}$$

$$\mathbf{P}(A|T = 2) = \frac{\mathbf{P}(\{T = 2\} \cap A)}{p_T(2)} = \frac{p_{T|A}(2) \mathbf{P}(A)}{p_T(2)} = \frac{1}{2}$$

. For any discrete random variable X and any collections of disjoint events A_1, A_2, \dots, A_n satisfying both $\mathbf{P}(A_j) > 0$ for all $j \in \{1, 2, \dots, n\}$ and $\sum_{i=1}^n \mathbf{P}(A_j) = 1$, we have

$$p_X(x) = \sum_{j=1}^n \mathbf{P}(A_j) p_{X|A_j}(x) \quad \forall x \quad \text{Total probability theorem, (6.29)}$$

$$\mathbf{P}(A_j|X = x) = \frac{\mathbf{P}(A_j) p_{X|A_j}(x)}{p_X(x)} \quad \forall x : p_X(x) > 0 \quad \text{Bayes' rule, (6.30)}$$

$$\mathbf{E}[X] = \sum_{j=1}^n \mathbf{P}(A_j) \mathbf{E}[X|A_j] \quad \text{Total expectation theorem. (6.31)}$$

On one hand (6.29) and (6.30) are mere restatements of (3.4) and (3.7) in terms of pmfs. On the other hand since they hold for all x values, they collectively imply equality of functions rather than numbers, e.g., (6.29) can be restated as $p_X = \sum_{j=1}^n \mathbf{P}(A_j) p_{X|A_j}$.

$$\begin{aligned} \mathbf{E}[X] &= \sum_x x p_X(x) \\ &= \sum_x \sum_{j=1}^n x \mathbf{P}(A_j) p_{X|A_j}(x) \quad \text{by (6.29)} \\ &= \sum_{j=1}^n \sum_x x \mathbf{P}(A_j) p_{X|A_j}(x) \quad \text{the order of the summations does not matter (why?)} \\ &= \sum_{j=1}^n \mathbf{P}(A_j) \mathbf{E}[X|A_j] \quad \text{by the definition of conditional expectation, (6.28).} \end{aligned}$$

HW. For the framework assumed for (6.29) and (6.31) express $p_{X|B}$ and $\mathbf{E}[X|B]$ in terms of $\mathbf{P}(A_j|B)$'s for an event B satisfying $\mathbf{P}(B) > 0$.

Example 6.20 (Memorylessness and Moments of Geometric Random Variables). Let X be a geometric random variable with success probability $p \in (0, 1)$ and A be the event that X takes a value strictly larger than $k \in \mathbb{Z}_+$, i.e.,

$$p_X(x) = \begin{cases} p(1-p)^{x-1} & \text{if } x = \{1, 2, \dots\} \\ 0 & \text{else} \end{cases} \quad A = \{X > k\}.$$

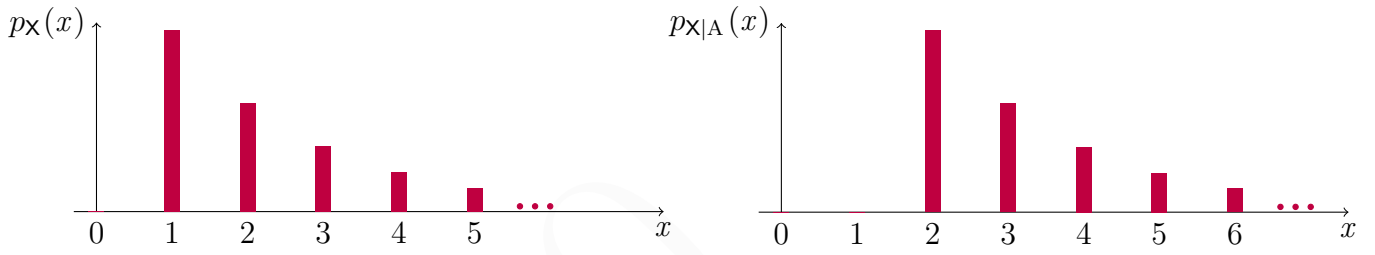
Determine $p_{X|A}$, $\mathbf{E}[X]$, and $\text{var}(X)$.

Let us start by calculating the probability of A

$$\begin{aligned} \mathbf{P}(A) &= \mathbf{P}(X > k) &&= \sum_{x>k} p_X(x) \\ &= (1-p)^k \sum_{y=1}^{\infty} p(1-p)^{y-1} &&= (1-p)^k \\ p_{X|A}(x) &= \mathbf{P}(X = x|A) \\ &= \frac{\mathbf{P}(\{X = x\} \cap A)}{\mathbf{P}(A)} \\ &= \begin{cases} \frac{\mathbf{P}(X=x)}{\mathbf{P}(A)} & x > k \\ 0 & x \leq k \end{cases} \\ &= \begin{cases} p(1-p)^{(x-k)-1} & x \in \{k+1, k+2, \dots\} \\ 0 & \text{else} \end{cases} \\ &= p_X(x-k) \\ &= p_Y(x) && \text{where } Y = X + k \end{aligned}$$

The conditional pmf of the geometric random variable X is just k units shifted version of the unconditional pmf, i.e., $p_{X|A}(x) = p_Y(x)$ where $Y = X + k$.

We only consider the event A for the case $k = 1$, i.e., $A = \{X > 1\}$, in the rest of our analysis.



$$\begin{aligned} \mathbf{E}[X] &= \mathbf{P}(A) \mathbf{E}[X|A] + \mathbf{P}(A^c) \mathbf{E}[X|A^c] \\ &= (1 - p)\mathbf{E}[X|A] + p\mathbf{E}[X|A^c] \\ &= (1 - p)\mathbf{E}[X|A] + p \\ &= (1 - p)\mathbf{E}[X + 1] + p \end{aligned}$$

by the total expectation theorem
 because $\mathbf{P}(A) = 1 - p$,
 because $p_{X|A^c}(1) = 1$,
 because $p_{X|A}(x) = p_Y(x)$ for $Y = X + 1$.

$$\mathbf{E}[X] = \frac{1}{p}$$

$$\begin{aligned} \mathbf{E}[X^2] &= \mathbf{P}(A) \mathbf{E}[X^2|A] + \mathbf{P}(A^c) \mathbf{E}[X^2|A^c] \\ &= (1 - p)\mathbf{E}[X^2|A] + p \\ &= (1 - p)\mathbf{E}[(X + 1)^2] + p \\ &= (1 - p)(\mathbf{E}[X^2] + 2\mathbf{E}[X] + 1) + p \end{aligned}$$

by the total expectation theorem
 by $\mathbf{P}(A) = 1 - p$ and $p_{X|A^c}(1) = 1$,
 because $p_{X|A}(x) = p_Y(x)$ for $Y = X + 1$,
 by the linearity of expectation,

$$\mathbf{E}[X^2] = \frac{2-p}{p^2}$$

$$\begin{aligned} \text{var}(X) &= \mathbf{E}[X^2] - (\mathbf{E}[X])^2 \\ &= \frac{1-p}{p^2} \end{aligned}$$

HW. Calculate the third moment of a geometric random variable with success probability p .

6.6.2 Conditional PMFs Given A Discrete Random Variable

Recall that any discrete random variable Y specifies a countable set of disjoint events of the form $\{Y = y\}$ with positive probability. For any discrete random variable X the conditional pmf $p_{X|Y}$ is defined using the conditional pmfs of the form $p_{X|\{Y=y\}}$ for such events as follows

$$\begin{aligned} p_{X|Y}(x|y) &:= \mathbf{P}(X = x|Y = y) && \forall y : \mathbf{P}(Y = y) > 0 && (6.32) \\ &= \frac{\mathbf{P}(X = x, Y = y)}{\mathbf{P}(Y = y)} && \forall y : \mathbf{P}(Y = y) > 0 && \text{by def. of conditional probability} \end{aligned}$$

$$= \frac{p_{X,Y}(x, y)}{p_Y(y)} \quad \forall y : p_Y(y) > 0 \quad \text{by def. joint pmf} \quad (6.33)$$

As expected for all values of y that $p_{X|Y}(\cdot|y)$ is defined it is a pmf, i.e.,

$$\sum_x p_{X|Y}(x|y) = 1 \quad \text{and} \quad p_{X|Y}(x|y) \geq 0 \quad \forall x.$$

In the following when conditioning on a discrete random variable, (i.e., conditioning on events of the form $\{Y = y\}$) the fact that this can be done only for positive values of the pmf $p_Y(y)$ (i.e., only for those y 's satisfying $\mathbf{P}(Y = y) > 0$), will be implicit.

We denote the conditional expectation of X given the event $\{Y = y\}$ by $\mathbf{E}[X|Y = y]$, i.e.,

$$\mathbf{E}[X|Y = y] := \sum_x x p_{X|Y}(x|y). \quad (6.34)$$

Conditioning on a discrete random variable Y is essentially conditioning on an event of the form $\{Y = y\}$ for different values of y : the conditional pmf $p_{X|Y}$ specifies all of the conditional pmfs $p_{X|Y=y}$ for values of y satisfying $p_Y(y) > 0$, i.e., satisfying $\mathbf{P}(Y = y) > 0$. Hence as one would expect the total probability theorem (6.29), Bayes' rule (6.30), and the total expectation (6.31) we have discussed for $p_{X|A}$ have their counter parts for $p_{X|Y}$.

. For discrete random variables X and Y ,

$$p_X(x) = \sum_y p_Y(y) p_{X|Y}(x|y) \quad \forall x \quad \text{Total probability th., (6.35)}$$

$$p_{Y|X}(y|x) = \frac{p_Y(y) p_{X|Y}(x|y)}{p_X(x)} \quad \forall x, y : p_{X,Y}(x, y) > 0 \quad \text{Bayes' rule, (6.36)}$$

$$\mathbf{E}[X] = \sum_y p_Y(y) \mathbf{E}[X|Y = y] \quad \text{Total expectation th. (6.37)}$$

As it was the case for conditional pmfs of the form $p_{X|A}$, for conditional pmfs of the form $p_{X|Y}$ as well, both the total probability theorem given in (6.35) and Bayes's rule given in (6.36) imply equality of functions.

Example 6.21. Let the the pmf of the random variable T and the conditional pmf of the random variable M given T be

$$p_T(t) = \begin{cases} \frac{3-|t-1|}{8} & \text{if } t \in \{0, 1, 2, 3\} \\ 0 & \text{else} \end{cases} \quad p_{M|T}(m|t) = \begin{cases} \frac{1}{4-t} & m \in \{t+1, \dots, 4\} \\ 0 & \text{else} \end{cases}$$

Calculate $\mathbf{E}[M]$ and $p_{T|M}(t|m)$

First note that $p_{M|T}(m|t)$ is the discrete uniform distribution on $t+1, \dots, 4$. Thus

$$\mathbf{E}[M|T = t] = \sum_m m p_{M|T}(m|t) = \frac{5+t}{2}$$

Then using the total expectation theorem we get

$$\mathbf{E}[M] = \sum_t p_T(t) \mathbf{E}[M|T = t] = \frac{2}{8} \cdot \frac{5}{2} + \frac{3}{8} \cdot \frac{6}{2} + \frac{2}{8} \cdot \frac{7}{2} + \frac{1}{8} \cdot \frac{8}{2} = \frac{25}{8}$$

We can use the indicator function $\mathbf{1}_{\{ \cdot \}}$ to express the pmfs p_T and $p_{M|T}(m|t)$ succinctly

$$p_T(t) = \frac{3-|t-1|}{8} \mathbf{1}_{\{t \in \{0,1,2,3\}\}} \quad p_{M|T}(m|t) = \frac{1}{4-t} \mathbf{1}_{\{m \in \{t+1, \dots, 4\}\}}$$

Then $p_M(m)$ can be positive only for $m \in \{1, 2, 3, 4\}$. Thus using Bayes' rule we have

$$p_{T|M}(t|m) = \frac{p_T(t) p_{M|T}(m|t)}{p_M(m)} = \frac{3-|t-1|}{p_M(m) \cdot 8 \cdot (4-t)} \mathbf{1}_{\{t \in \{0, \dots, m-1\}\}}$$

We calculate $p_{T|M}(\cdot|m)$ using the fact that $\sum_t p_{T|M}(t|m) = 1$ for all m for which it is defined.

$$\begin{aligned} p_{T|M}(t|1) &= \mathbf{1}_{\{t=0\}} & p_{T|M}(t|2) &= \frac{1}{3} \mathbf{1}_{\{t=0\}} + \frac{2}{3} \mathbf{1}_{\{t=1\}} \\ p_{T|M}(t|3) &= \frac{1}{5} \mathbf{1}_{\{t=0\}} + \frac{2}{5} \mathbf{1}_{\{t \in \{1,2\}\}} & p_{T|M}(t|4) &= \frac{1}{7} \mathbf{1}_{\{t=0\}} + \frac{2}{7} \mathbf{1}_{\{t \in \{1,2,3\}\}} \end{aligned}$$

HW. Confirm that the joint distribution $p_{T,M}$ is the one obtained in Example 6.15

Example 6.22. Attendees of a party are required to have a hat; n attendees of this party throw their hats into a bag and then pick one hat each from the bag. All assignment of n hats to n attendees are equally likely and X is the number of attendees who get their own hat back. Determine $\text{var}(X)$.

$$X = \sum_{i=1}^n X_i \quad \text{where} \quad X_i = \begin{cases} 1 & \text{if } i^{\text{th}} \text{ individual gets their hat back} \\ 0 & \text{else} \end{cases}$$

$$\text{var}(X) = \mathbf{E}[X^2] - (\mathbf{E}[X])^2$$

Let us proceed with calculating $\mathbf{E}[X]$ and $\mathbf{E}[X^2]$,

$$\begin{aligned} \mathbf{E}[X] &= \mathbf{E}\left[\sum_{i=1}^n X_i\right] \\ &= \sum_{i=1}^n \mathbf{E}[X_i] && \text{by the linearity of the expectation} \\ &= 1 && \text{by } \mathbf{E}[X_i] = p_{X_i}(1) = n^{-1} \text{ for all } i \end{aligned}$$

$$\begin{aligned} \mathbf{E}[X^2] &= \mathbf{E}\left[\left(\sum_{i=1}^n X_i\right)\left(\sum_{j=1}^n X_j\right)\right] \\ &= \mathbf{E}\left[\sum_{i=1}^n \sum_{j=1}^n X_i X_j\right] && \text{by the linearity of the expectation} \\ &= \sum_{i=1}^n \sum_{j=1}^n \mathbf{E}[X_i X_j] && \text{for } i \neq j \\ &= \sum_{i=1}^n \mathbf{E}[X_i^2] + \sum_{i=1}^n \sum_{j \neq i} \mathbf{E}[X_i X_j] && \text{by } X_i^2 = X_i \text{ and } \mathbf{E}[X_i] = n^{-1} \\ &= 1 + \sum_{i=1}^n \sum_{j \neq i} \mathbf{E}[X_i X_j] \end{aligned}$$

$$\begin{aligned} \mathbf{E}[X_i X_j] &= p_{X_i}(0) \mathbf{E}[X_i X_j | X_i = 0] + p_{X_i}(1) \mathbf{E}[X_i X_j | X_i = 1] \\ &= n^{-1} \mathbf{E}[X_j | X_i = 1] && \text{by } \mathbf{E}[X_i X_j | X_i = 0] = 0 \text{ and } p_{X_i}(1) = n^{-1}, \\ &= (n(n-1))^{-1} && \text{by } \mathbf{E}[X_j | X_i = 1] = p_{X_j | X_i}(1|1) = (n-1)^{-1} \end{aligned}$$

$$\mathbf{E}[X^2] = 2$$

$$\text{var}(X) = 1$$

The conditional probability law can also be used to determine the “conditional” joint pmf for a finite collection of discrete random variables, X_1, \dots, X_n just as it is used in (6.27) to calculate the conditional pmf for a single random variable:

$$p_{X_1, \dots, X_n | A}(x_1, \dots, x_n) := \mathbf{P}(X_1 = x_1, \dots, X_n = x_n | A) \quad \forall x_1, \dots, x_n. \quad (6.38)$$

Furthermore, instead of conditioning on events of the form $\{Y = y\}$ about a single random variable as we did in (6.32), one can condition on events of the form $\{Y_1 = y_1, \dots, Y_k = y_k\}$

$$p_{X_1, \dots, X_n | Y_1, \dots, Y_k}(x_1, \dots, x_n | y_1, \dots, y_k) := \mathbf{P}(X_1 = x_1, \dots, X_n = x_n | Y_1 = y_1, \dots, Y_n = y_n). \quad (6.39)$$

The following short hand $Z_k^n = (Z_k, \dots, Z_n)$ is useful when working with such relations

$$\begin{aligned} p_{X_1^n | A}(x_1^n) &:= \mathbf{P}(X_1^n = x_1^n | A) && \forall x_1^n \\ p_{X_1^n | Y_1^k}(x_1^n | y_1^k) &:= \mathbf{P}(X_1^n = x_1^n | Y_1^k = y_1^k) && \forall y : \mathbf{P}(Y_1^k = y_1^k) > 0 \\ &= \frac{p_{X_1^n, Y_1^k}(x_1^n, y_1^k)}{p_{Y_1^k}(y_1^k)} && \forall y_1^k : p_{Y_1^k}(y_1^k) > 0 \end{aligned}$$

HW. Let X_1, \dots, X_n be discrete random variables and $\ell \in \{1, \dots, n-1\}$. State the total probability theorem to express $p_{X_{\ell+1}^n}$ in terms of $p_{X_{\ell+1}^n | X_1^\ell}$. State the Bayes’ law to express $p_{X_1^\ell | X_{\ell+1}^n}$ in terms of $p_{X_{\ell+1}^n | X_1^\ell}$.

6.7 Conditional Expectation and Iterated Expectations

Let X and Y be a discrete random variables and X has a finite expectation. Then as a result of the total expectation theorem given in (6.37) the conditional expectation $\mathbf{E}[X|Y = y]$ defined in (6.34) satisfy the following relation

$$\mathbf{E}[X] = \sum_y p_Y(y) \mathbf{E}[X|Y = y]. \tag{6.40}$$

Thus $\mathbf{E}[X|Y = y]$ maps each y satisfying $p_Y(y) > 0$ to a real number, hence the conditional expectation of X given Y effectively defines a function. Let us denote this function by $g(\cdot)$, i.e.,

$$g(y) = \mathbf{E}[X|Y = y] \qquad \forall y : p_Y(y) > 0. \tag{6.41}$$

We know that $g(Y)$ is a discrete random variable because any function of a discrete random variable is. Furthermore we know $\mathbf{E}[g(Y)] = \mathbf{E}[X]$ as a result of (6.40) and (6.41).

To emphasize the fact that conditional expectation of X given Y is a random variable, it is denoted by $\mathbf{E}[X|Y]$ and the total expectation theorem given in (6.37) and restated in (6.40) is stated as follows

$$\mathbf{E}[X] = \mathbf{E}[\mathbf{E}[X|Y]] \qquad \text{The Law of Iterated Expectations.} \tag{6.42}$$

Recall that we have assumed that $\mathbf{E}[X]$ to be a real number. Then $\mathbf{E}[X|Y = y]$ is real number for all y that it is defined for by (6.40) and $\mathbf{E}[X|Y]$ is a random variable whose value is determined by Y . If $\mathbf{E}[X]$ is undefined or infinite ($-\infty$ or ∞), then the conditional expectation $\mathbf{E}[X|Y = y]$ may or may not be a real number a y satisfying $p_Y(y) > 0$. Thus we maynot assert that $\mathbf{E}[X|Y]$ is a random variable, unless $\mathbf{E}[X]$ is defined and finite.

Example 6.23. The random variable Y has a geometric distribution with success probability p , the conditional pmfs of the random variable X given Z are

$$p_{Z|Y}(z|y) = \frac{1}{2} \mathbf{1}_{\{z=(1-p)^{-y}\}} + \frac{1}{2} \mathbf{1}_{\{z=-(1-p)^{-y}\}} \qquad p_{X|Y}(x|y) = e^{-y} \frac{y^x}{x!} \mathbf{1}_{\{x \in \mathbb{Z}_{\geq 0}\}}$$

Determine $\mathbf{E}[Z|Y]$, $\mathbf{E}[Z]$, $\mathbf{E}[X]$, $\mathbf{var}(X)$, $\mathbf{E}[XY]$.

$\mathbf{E}[Z|Y] = 0$

$\mathbf{E}[Z]$ is undefined because $\mathbf{E}[Z^+] = \infty$ and $\mathbf{E}[Z^-] = \infty$.

$\mathbf{E}[X] = \mathbf{E}[\mathbf{E}[X|Y]]$ by the law of iterated expectations (note that $\mathbf{P}(X \geq 0) = 1$)

$= \mathbf{E}[Y]$ the expectation of a Poisson r.v. with parameter λ is λ

$= \frac{1}{p}$ the expectation of a Geometric r.v. with parameter p is $\frac{1}{p}$

$\mathbf{E}[X^2] = \mathbf{E}[\mathbf{E}[X^2|Y]]$ by the law of iterated expectations

$= \mathbf{E}[Y + Y^2]$ the second moment of a Poisson r.v. with parameter λ is $\lambda + \lambda^2$

$= \frac{1}{p} + \frac{2-p}{p^2}$ the first two moments of a Geometric r.v. are $\frac{1}{p}$ and $\frac{2-p}{p^2}$

$\mathbf{var}(X) = \frac{1}{p^2}$ because $\mathbf{var}(X) = \mathbf{E}[X^2] - (\mathbf{E}[X])^2$

$\mathbf{E}[XY] = \mathbf{E}[\mathbf{E}[XY|Y]]$ by the law of iterated expectations

$= \mathbf{E}[Y\mathbf{E}[X|Y]]$ because $\mathbf{E}[YX|Y] = Y\mathbf{E}[X|Y]$ via the interpretation of $\mathbf{E}[X|Y]$ as $g(Y)$

$= \mathbf{E}[Y^2]$ the expectation of a Poisson r.v. is λ , i.e., $\mathbf{E}[X|Y] = Y$,

$= \frac{2-p}{p^2}$ by the second moment of a Geometric r.v. with parameter p

An important consequence of interpretation of conditional expectation as a random variable is that the conditional expectation of the product of a random variable and a function of the conditioned random variable is equal to the product of the conditional expectation and the function of the random variable, i.e., let X and Y be discrete random variables and h be a real valued function then

$$\mathbf{E}[h(Y)X|Y] = h(Y)\mathbf{E}[X|Y]. \tag{6.43}$$

Note that (6.43) asserts the equality of two random variables; not just two numbers.

6.7.1 Conditional Variance and Law of Total Variance

The conditional variance of X given Y is the variance of the random variable X given the value of the random variable Y :

$$\mathbf{var}(X|Y = y) := \sum_x (x - \mathbf{E}[X|Y = y])^2 p_{X|Y}(x|y). \tag{6.44}$$

$$\mathbf{var}(X|Y) := \mathbf{E}[(X - \mathbf{E}[X|Y])^2|Y] \tag{6.45}$$

Recall that the variance of a discrete random variable can be expressed as the difference of the second moment and the square of the first moment, see (6.14). The corresponding relation can be derived for the conditional variance as well.

$$\begin{aligned} \mathbf{var}(X|Y) &= \mathbf{E}[(X - \mathbf{E}[X|Y])^2|Y] \\ &= \mathbf{E}[X^2 - 2X\mathbf{E}[X|Y] + (\mathbf{E}[X|Y])^2|Y] \\ &= \mathbf{E}[X^2|Y] - 2\mathbf{E}[X\mathbf{E}[X|Y]|Y] + \mathbf{E}[(\mathbf{E}[X|Y])^2|Y] \quad \text{by the linearity of the expectation} \\ &= \mathbf{E}[X^2|Y] - 2\mathbf{E}[X|Y]\mathbf{E}[X|Y] + (\mathbf{E}[X|Y])^2 \quad \text{by (6.43) for } h(Y) = \mathbf{E}[X|Y]. \\ &= \mathbf{E}[X^2|Y] - (\mathbf{E}[X|Y])^2 \end{aligned} \tag{6.46}$$

The variance of a random variable can be expressed in terms of its conditional variance. Inspired by the names used for similar relations for probability and expectation, corresponding relation is called the total variance theorem. However, unlike relations with the similar name variance is not just average of conditional variance, it has a an additional term associated with variance of the conditional expectation.

$$\mathbf{var}(X) = \mathbf{E}[\mathbf{var}(X|Y)] + \mathbf{var}(\mathbf{E}[X|Y]) \quad \text{Total Variance Theorem} \tag{6.47}$$

$$\begin{aligned} \mathbf{var}(X) &= \mathbf{E}[X^2] - (\mathbf{E}[X])^2 && \text{by (6.14),} \\ &= \mathbf{E}[\mathbf{E}[X^2|Y]] - (\mathbf{E}[X])^2 && \text{by the law of iterated expectations,} \\ &= \mathbf{E}[\mathbf{var}(X|Y) + (\mathbf{E}[X|Y])^2] - (\mathbf{E}[X])^2 && \text{by (6.46),} \\ &= \mathbf{E}[\mathbf{var}(X|Y)] + \mathbf{E}[(\mathbf{E}[X|Y])^2] - (\mathbf{E}[X])^2 && \text{by the linearity of the expectaion,} \\ &= \mathbf{E}[\mathbf{var}(X|Y)] + \mathbf{E}[(\mathbf{E}[X|Y])^2] - (\mathbf{E}[\mathbf{E}[X|Y]])^2 && \text{by the law of iterated expectations,} \\ &= \mathbf{E}[\mathbf{var}(X|Y)] + \mathbf{var}(\mathbf{E}[X|Y]) && \text{by (6.14).} \end{aligned}$$

HW. Prove the law of total variance using the definition of variance and (6.43), without explicitly relying on (6.14) or (6.46). Hint: $(X - \mathbf{E}[X])^2 = [(X - \mathbf{E}[X|Y]) + (\mathbf{E}[X|Y] - \mathbf{E}[X])]^2$

For the random variable \mathbf{X} considered in Example 6.23

$$\begin{aligned} \mathbf{var}(\mathbf{X}|\mathbf{Y}) &= \mathbf{Y} && \text{because } p_{\mathbf{X}|\mathbf{Y}}(\cdot|y) \text{ is Poisson with parameter } y. \\ \mathbf{var}(\mathbf{E}[\mathbf{X}|\mathbf{Y}]) &= \mathbf{var}(\mathbf{Y}) && \text{because } p_{\mathbf{X}|\mathbf{Y}}(\cdot|y) \text{ is Poisson with parameter } y. \\ \mathbf{var}(\mathbf{X}) &= \mathbf{E}[\mathbf{Y}] + \mathbf{var}(\mathbf{Y}) \\ &= \frac{1}{p^2} && \text{because } p_{\mathbf{Y}} \text{ is Geometric and hence } \mathbf{E}[\mathbf{Y}] = \frac{1}{p} \text{ and } \mathbf{var}(\mathbf{Y}) = \frac{1-p}{p^2}. \end{aligned}$$

Example 6.24 (A variant of the two envelope paradox [BT08, Example 2.18]). Let \mathbf{X} be a Geometric rv with success probability p , i.e. $p_{\mathbf{X}}(k) = p(1-p)^{k-1}\mathbb{1}_{\{k \in \mathbb{Z}_+\}}$. If $\mathbf{X} = k$, then Alice puts m^k liras in the envelope \mathbf{Y} and m^{k+1} liras in the envelope \mathbf{Z} with probability $1/2$ and m^{k+1} liras in the envelope \mathbf{Y} and m^k liras in the envelope \mathbf{Z} with probability $1/2$, where m is an integer strictly greater than 1, i.e., $m \in \{2, 3, \dots\}$. You know the values of parameters p and m . You open the envelope \mathbf{Y} and observe that it has y liras in it. You can either keep y liras or switch envelopes and get the money in the envelope \mathbf{Z} instead. If you want to maximize your expected earnings from this game, for which values of y will you switch envelopes? Is there a value of p and m for which you will always switch envelopes?

Solution. With a slight abuse of notation let us denote the amount money—in liras—in the envelope \mathbf{Y} by \mathbf{Y} and in the envelope \mathbf{Z} by \mathbf{Z} . The expected amount of money we get by switching to the envelope \mathbf{Z} is $\mathbf{E}[\mathbf{Z}|\mathbf{Y}]$. To maximize our earnings we need to switch whenever $\mathbf{E}[\mathbf{Z}|\mathbf{Y}] > \mathbf{Y}$.

Remark 6.5. Whenever $\mathbf{E}[\mathbf{Z}|\mathbf{Y}] = \mathbf{Y}$ holds, switching to the envelope \mathbf{Z} will not change the expected earnings. If we switch to the envelope \mathbf{Z} whenever $\mathbf{E}[\mathbf{Z}|\mathbf{Y}] \geq \mathbf{Y}$, the expected gain will be the same. Thus our rule for switching to the envelope \mathbf{Z} can be $\mathbf{E}[\mathbf{Z}|\mathbf{Y}] > \mathbf{Y}$ or $\mathbf{E}[\mathbf{Z}|\mathbf{Y}] \geq \mathbf{Y}$.

In order to determine the values of y satisfying $\mathbf{E}[\mathbf{Z}|\mathbf{Y} = y] > y$, we need to calculate the expected value of \mathbf{Z} given \mathbf{Y} and hence the conditional PMF of \mathbf{Z} given \mathbf{Y} . Note that $p_{\mathbf{Y},\mathbf{Z}|\mathbf{X}}$ is described in the question and $p_{\mathbf{X}}$ is given. Then

$$\begin{aligned} p_{\mathbf{Y},\mathbf{Z}|\mathbf{X}}(y, z|k) &= \frac{1}{2}\mathbb{1}_{\{(y,z) \in \{(m^k, m^{k+1}), (m^{k+1}, m^k)\}\}} && (6.48) \\ p_{\mathbf{Y},\mathbf{Z}}(y, z) &= \frac{p(1-p)^{k-1}}{2}\mathbb{1}_{\{(y,z) \in \{(m^k, m^{k+1}), (m^{k+1}, m^k)\}, k \in \mathbb{Z}_+\}} && \text{by } p_{\mathbf{Y},\mathbf{Z}}(\cdot) = \sum_k p_{\mathbf{X}}(k) p_{\mathbf{Y},\mathbf{Z}|\mathbf{X}}(\cdot|k) \\ p_{\mathbf{Y}}(y) &= \frac{p}{2}\mathbb{1}_{\{y=m\}} + \frac{(2-p) \cdot p \cdot (1-p)^{k-2}}{2}\mathbb{1}_{\{y=m^k, k \in \mathbb{Z}_{\geq 2}\}} && \text{by } p_{\mathbf{Y}}(y) = \sum_z p_{\mathbf{Y},\mathbf{Z}}(y, z) \end{aligned}$$

Then using $p_{\mathbf{Z}|\mathbf{Y}} = \frac{p_{\mathbf{Y},\mathbf{Z}}}{p_{\mathbf{Y}}}$ and $\mathbf{E}[\mathbf{Z}|\mathbf{Y} = y] = \sum_z z p_{\mathbf{Z}|\mathbf{Y}}(z|y)$ we get

$$\begin{aligned} p_{\mathbf{Z}|\mathbf{Y}}(z|m) &= \mathbb{1}_{\{z=m^2\}} && p_{\mathbf{Z}|\mathbf{Y}}(z|m^k) = \frac{1-p}{2-p}\mathbb{1}_{\{z=m^{k+1}\}} + \frac{1}{2-p}\mathbb{1}_{\{z=m^{k-1}\}} \\ \mathbf{E}[\mathbf{Z}|\mathbf{Y} = m] &= m^2 && \mathbf{E}[\mathbf{Z}|\mathbf{Y} = m^k] = m^{k+1}\frac{1-p}{2-p} + m^{k-1}\frac{1}{2-p} \\ &&& = m^k \left(1 + \frac{m-1}{2-p} \left(1 - p - \frac{1}{m} \right) \right) \end{aligned}$$

If $y = m$ we switch to the envelope \mathbf{Z} , because $\mathbf{E}[\mathbf{Z}|\mathbf{Y} = m] = m^2 > m$. In fact we know that it has exactly m^2 liras with probability 1.

If $y = m^k$ for some $k \in \{2, 3, \dots\}$, then $\mathbf{E}[\mathbf{Z}|\mathbf{Y} = y] > y$ holds if and only if $m > \frac{1}{1-p}$. Thus

- If $m > \frac{1}{1-p}$, then $\mathbf{E}[\mathbf{Z}|\mathbf{Y} = m^k] > m^k$ and we switch to envelope \mathbf{Z} for all $k \in \{2, 3, \dots\}$.
- If $m = \frac{1}{1-p}$, then $\mathbf{E}[\mathbf{Z}|\mathbf{Y} = m^k] = m^k$ we may switch to \mathbf{Z} or retain \mathbf{Y} for all $k \in \{2, 3, \dots\}$.
- If $m < \frac{1}{1-p}$, then $\mathbf{E}[\mathbf{Z}|\mathbf{Y} = m^k] < m^k$ we retain envelope \mathbf{Y} for all $k \in \{2, 3, \dots\}$.

Thus a strategy that maximize the expected earnings is given as follows for different values of the parameters m and p :

If $m > \frac{1}{1-p}$, then we will always switch to the envelope Z .

If $1 < m \leq \frac{1}{1-p}$, then we will switch to the envelope Z if $y = 1$ and retain the envelope Y else.

Remark 6.6. $\mathbf{E}[Y|X] = \mathbf{E}[Z|X]$ by (6.48). Thus $\mathbf{E}[Y] = \mathbf{E}[Z]$ by the law of iterated expectations. Thus one might be tempted to doubt the conclusion that “If $m > \frac{1}{1-p}$, then we will always switch to the envelope Z .” Our conclusions, however, is sound; the subtlety here is that if $m > \frac{1}{1-p}$ then both $\mathbf{E}[Y]$ and $\mathbf{E}[Z]$ are infinite.

$$\begin{aligned}\mathbf{E}[Y] &= \sum_{k=1}^{\infty} p(1-p)^{k-1} \frac{m^k + m^{k+1}}{2} \\ &= \frac{(m+1)mp}{2} \sum_{k=1}^{\infty} ((1-p)m)^k\end{aligned}$$

6.8 Independence

6.8.1 Independence of a Random Variable From an Event

A discrete random variable X is independent of an event A iff

$$\begin{aligned} \mathbf{P}(\{X = x\} \cap A) &= \mathbf{P}(X = x) \mathbf{P}(A) && \forall x && (6.49) \\ &= p_X(x) \mathbf{P}(A) && \forall x \end{aligned}$$

If $\mathbf{P}(A) = 0$, then A is independent of any discrete random variable X . If $\mathbf{P}(A) > 0$, then for any discrete random variable X the conditional pmf $p_{X|A}$ is defined and X is independent of A iff

$$p_{X|A}(x) = p_X(x) \quad \forall x \quad (6.50)$$

Note that for a discrete random variable X there exist a countable set Θ such that $\mathbf{P}(X \notin \Theta) = 0$. Thus for any set of real numbers S we have

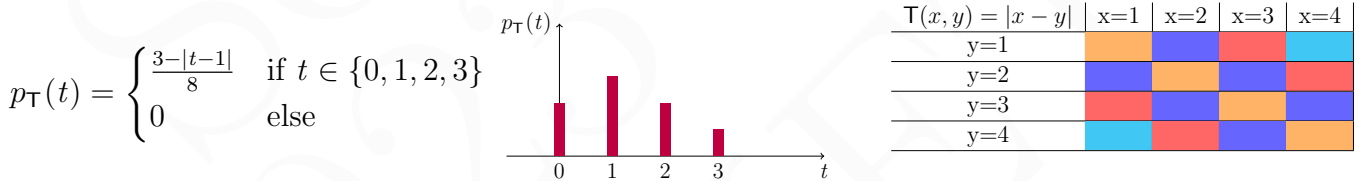
$$\begin{aligned} \mathbf{P}(\{X \in S\} \cap A) &= \mathbf{P}(\{X \in S \cap \Theta\} \cap A) + \mathbf{P}(\{X \in S \setminus \Theta\} \cap A) && \text{by the additivity of probability} \\ &= \mathbf{P}(\{X \in S \cap \Theta\} \cap A) && \mathbf{P}(\{X \in S \setminus \Theta\}) = 0 \\ &= \sum_{x \in S \cap \Theta} \mathbf{P}(\{X = x\} \cap A) && \text{by the } \sigma\text{-additivity of probability} \end{aligned}$$

Then as a result of (6.1) the constraint in (6.49) holds iff all events about the random variable X , (i.e., all events of the form $\{X \in S\}$ for some set of real numbers S) are independent of the event A .

Example 6.25. Let us consider two independent rolls of a fair tetrahedral die. Let the event A , and the random variables T and M be

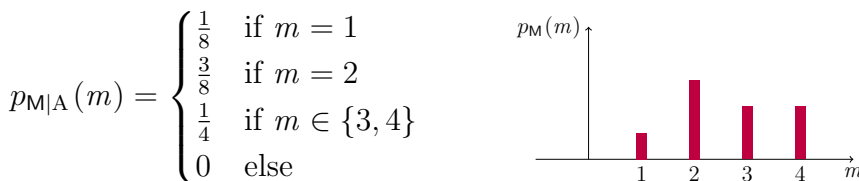
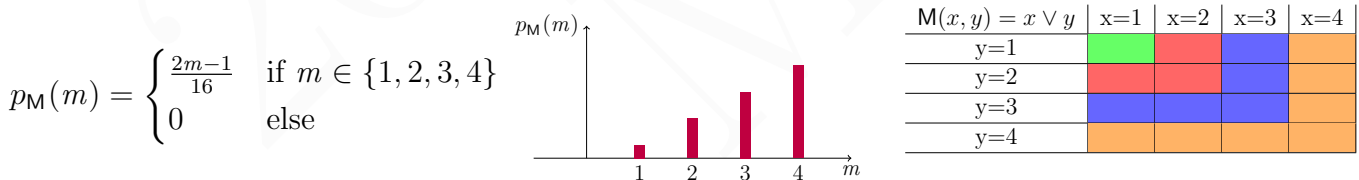
$$A = \{(x, y) : x \in \{1, 2\}\} \quad T(x, y) = |x - y| \quad M(x, y) = x \vee y$$

Is the random variable T independent of the event A ?



Note that $p_T = p_{T|A}$ thus T is independent of A .

Is the random variable M independent of the event A ?



The random variable M is not independent of the event A because $p_{M|A} \neq p_M$.

6.8.2 Independence of Two Random Variables

Two discrete random variables X and Y are independent iff

$$p_{X,Y}(x, y) = p_X(x) p_Y(y) \quad \forall x, y. \tag{6.51}$$

That is discrete random variables X and Y are independent iff events $\{X = x\}$ and $\{Y = y\}$ for all x and y . If $p_Y(y) = 0$, then the events $\{X = x\}$ and $\{Y = y\}$ are independent for any x . If $p_Y(y) > 0$, then the conditional pmf $p_{X|Y}(\cdot|y)$ is defined and the constraint given in (6.51) is satisfied iff $p_{X|Y}(\cdot|y) = p_X(\cdot)$. Thus the constraint given in (6.51) is equivalent to

$$p_{X|Y}(x|y) = p_X(x) \quad \forall x, y : p_Y(y) > 0. \tag{6.52}$$

Pursuing an analysis similar to the one for independence of a random variable from an event, one can show that discrete random variables X and Y are independent iff any event about X is independent of any event about Y , i.e. discrete random variables X and Y are independent iff

$$\mathbf{P}(X \in S, Y \in T) = \mathbf{P}(X \in S) \mathbf{P}(Y \in T) \quad \forall S \subset \mathbb{R}, T \subset \mathbb{R} \tag{6.53}$$

HW. Prove that two discrete random variables X and Y are independent iff the constraint given (6.53) is satisfied. *Hint:* Follow the argument for independence of a random variable from an event via the countable sets Θ_1 and Θ_2 satisfying $\mathbf{P}(X \notin \Theta_1) = 0$ and $\mathbf{P}(Y \notin \Theta_2) = 0$.

As we did for events we can define conditional independence for random variables. Let A be an event satisfying $\mathbf{P}(A) > 0$ and X and Y be discrete random variables; then X and Y are said to be conditionally independent given event A iff

$$p_{X,Y|A}(x, y) = p_{X|A}(x) p_{Y|A}(y) \quad \forall x, y. \tag{6.54}$$

Equivalently, X and Y are said to be conditionally independent given event A iff

$$p_{X|Y,A}(x|y) = p_{X|A}(x) \quad \forall x, y : p_{Y|A}(y) > 0. \tag{6.55}$$

As it was the case for events independence of two random variables does not necessarily imply their conditional independence given an event with positive probability. Furthermore, conditional independence of two random variables does not imply their independence.

Example 6.26. Let X and Y be random variable describing the outcome two independent rolls of a fair four faced die and the event A be $\{X \geq Y\}$. Are X and Y conditional independent given A ?

$$p_{X,Y}(x, y) = p_X(x) p_Y(y)$$

$$p_{X|A}(x) = \begin{cases} 1/4 & \text{if } x \in \{1, 2, 3, 4\} \\ 0 & \text{else} \end{cases} \quad p_Y(y) = p_X(y)$$

$$\mathbf{P}(X \geq Y) = 10/16$$

$y \backslash x$	1	2	3	4
1	1/16	1/16	1/16	1/16
2	1/16	1/16	1/16	1/16
3	1/16	1/16	1/16	1/16
4	1/16	1/16	1/16	1/16

$$p_{X,Y|A}(x, y) = \begin{cases} 1/10 & \text{if } x \in \{1, 2, 3, 4\}, y \in \{1, \dots, x\} \\ 0 & \text{else} \end{cases}$$

$$p_{X|A}(x) = \begin{cases} x/10 & \text{if } x \in \{1, 2, 3, 4\} \\ 0 & \text{else} \end{cases}$$

$$p_{Y|A}(y) = \begin{cases} \frac{5-y}{10} & \text{if } y \in \{1, 2, 3, 4\} \\ 0 & \text{else} \end{cases}$$

$$p_{X,Y|A} \neq p_{X|A} p_{Y|A}$$

$y \backslash x$	1	2	3	4
1	1/10	1/10	1/10	1/10
2	0	1/10	1/10	1/10
3	0	0	1/10	1/10
4	0	0	0	1/10

Example 6.27. Let X and Y random variable with the following pmf and let the event A be $X + Y < 6$. Are X and Y conditional independent given A ? Are X and Y independent?

$x \backslash y$	1	2	3
1	$1/24$	$1/24$	0
2	$5/24$	$5/24$	0
3	$1/12$	$1/12$	$1/3$

$$p_X(x) = \begin{cases} 1/3 & \text{if } x \in \{1, 2, 3\} \\ 0 & \text{else} \end{cases} \quad p_Y(y) = \begin{cases} 1/12 & \text{if } y = 1 \\ 5/12 & \text{if } y = 2 \\ 6/12 & \text{if } y = 3 \\ 0 & \text{else} \end{cases}$$

$$p_{X,Y} \neq p_X p_Y$$

$$p_{X,Y|A} = p_{X|A} p_{Y|A} \quad p_{X|A}(x) = \begin{cases} 1/2 & \text{if } x \in \{1, 2\} \\ 0 & \text{else} \end{cases} \quad p_{Y|A}(y) = \begin{cases} 1/8 & \text{if } y = 1 \\ 5/8 & \text{if } y = 2 \\ 2/8 & \text{if } y = 3 \\ 0 & \text{else} \end{cases}$$

. If discrete random variables X and Y are independent, then

$$\mathbf{E}[XY] = \mathbf{E}[X] \mathbf{E}[Y] \tag{6.56}$$

$$\begin{aligned} \mathbf{E}[XY] &= \sum_x \sum_y xy p_{X,Y}(x, y) \\ &= \sum_x \sum_y xy p_X(x) p_Y(y) && \text{by the independence} \\ &= \sum_x x p_X(x) \sum_y y p_Y(y) \\ &= \mathbf{E}[X] \mathbf{E}[Y] \end{aligned}$$

. Let X and Y be independent discrete random variables then for any pair of function $g(\cdot)$ and $h(\cdot)$, discrete random variables $G = g(X)$ and $H = h(Y)$ are independent.

$$\begin{aligned} p_{G,H}(a, b) &= \mathbf{P}(G = a, H = b) \\ &= \mathbf{P}(g(X) = a, h(Y) = b) \\ &= \mathbf{P}(X \in \{x : g(x) = a\}, Y \in \{y : h(y) = b\}) \\ &= \mathbf{P}(X \in \{x : g(x) = a\}) \mathbf{P}(Y \in \{y : h(y) = b\}) && \text{by the independence of } X \text{ and } Y \\ &= \mathbf{P}(g(X) = a) \mathbf{P}(h(Y) = b) \\ &= p_G(a) p_H(b) \end{aligned}$$

For any pair of discrete random variables X and Y that are and any pair of functions $g(\cdot)$ and $h(\cdot)$, we have

$$\mathbf{E}[g(X)h(Y)] = \mathbf{E}[g(X)] \mathbf{E}[h(Y)] \tag{6.57}$$

HW. Prove (6.57) without explicitly relying on the independence of $g(X)$ and $h(Y)$.

. If discrete random variable X and Y are independent, then

$$\mathbf{var}(X + Y) = \mathbf{var}(X) + \mathbf{var}(Y) \tag{6.58}$$

$$\begin{aligned} \mathbf{var}(X + Y) &= \mathbf{E}[(X + Y - \mathbf{E}[X + Y])^2] && \text{by def. of variance} \\ &= \mathbf{E}[(X - \mathbf{E}[X] + Y - \mathbf{E}[Y])^2] && \text{by lin. of expectation} \\ &= \mathbf{var}(X) + 2\mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])] + \mathbf{var}(Y) && \text{by def. of variance} \\ &= \mathbf{var}(X) + 2\mathbf{E}[X - \mathbf{E}[X]] \mathbf{E}[Y - \mathbf{E}[Y]] + \mathbf{var}(Y) && \text{by independence, i.e. (6.57)} \\ &= \mathbf{var}(X) + \mathbf{var}(Y) && \text{because } \mathbf{E}[X - \mathbf{E}[X]] = 0 \end{aligned}$$

6.8.3 Independence of Several Random Variables

Discrete random variables X_1, X_2, \dots, X_n are independent iff

$$p_{X_1^n}(x_1^n) = \prod_{j=1}^n p_{X_j}(x_j) \quad \forall x_1^n \quad (6.59)$$

For the case of three random variables for example, we have the following. Discrete random variables X, Y , and Z are independent iff

$$p_{X,Y,Z}(x, y, z) = p_X(x) p_Y(y) p_Z(z) \quad \forall x, y, z.$$

Independence of random variables X_1, X_2, \dots, X_n imply the independence of any sub-collection of the set of random variables from X_1, X_2, \dots, X_n . In order to see why let us establish the independence of random variables X_1, X_2, \dots, X_k for a $1 < k < n$, for the case when random variables X_1, X_2, \dots, X_n are independent.

$$\begin{aligned} p_{X_1^k}(x_1^k) &= \sum_{x_{k+1}^n} p_{X_1^n}(x_1^n) && \text{by definition of the joint distributions} \\ &= \sum_{x_{k+1}^n} \prod_{j=1}^n p_{X_j}(x_j) && \text{by independence of } X_1, X_2, \dots, X_n \\ &= \left(\prod_{j=1}^k p_{X_j}(x_j) \right) \sum_{x_{k+1}^n} p_{X_{k+1}}(x_{k+1}) \cdots p_{X_n}(x_n) \\ &= \prod_{j=1}^k p_{X_j}(x_j) \end{aligned}$$

Recall that any sub-collection of an independent collection of events was also independent.

Independence several random variables can be characterized in terms of independence of certain events, similar to the independence of two random variables. Discrete random variables X_1, X_2, \dots, X_n are independent iff events $\{X_1 \in S_1\}, \dots, \{X_n \in S_n\}$ are independent for all $S_1 \subset \mathbb{R}, \dots, S_n \subset \mathbb{R}$.

Remark 6.7. For events A_1, \dots, A_n to be independent we required the following constraint

$$\mathbf{P}(\cap_{j \in S} A_j) = \prod_{j \in S} \mathbf{P}(A_j) \quad S \subset \{1, \dots, n\}. \quad (6.60)$$

We required the constraint to hold for all subsets S because we essentially wanted to assert the following rule

$$\mathbf{P}(\cap_{j=1}^n B_j) = \prod_{j=1}^n \mathbf{P}(B_j) \quad \text{where } B_j \in \{\emptyset, A, A^c, \Omega\} \text{ for all } j \in \{1, \dots, n\}. \quad (6.61)$$

One can show that not only that constraints in (6.60) and (6.61) are equivalent, but also they are equivalent to the independence of the random variables $\mathbb{1}_{\{A_1\}}, \dots, \mathbb{1}_{\{A_n\}}$.

If random variables X_1, X_2, \dots, X_n are independent then random variables obtain as a function of disjoint sub-collections of these random variable will be independent as well. For example random variable $g(X_1^k), h(X_{k+1}^\ell), f(X_{\ell+1}^n)$ will be independent. But the pair of random variables $g(X_1^k), h(X_k^n)$ will not necessarily be independent.

If discrete random variables X_1, X_2, \dots, X_n are independent, then

$$\begin{aligned} \mathbf{E}\left[\prod_{j=1}^n X_j\right] &= \sum_{x_1^n} \left(\prod_{j=1}^n x_j \right) p_{X_1^n}(x_1^n) \\ &= \sum_{x_1^n} \prod_{j=1}^n x_j p_{X_j}(x_j) && \text{by independence} \\ &= \sum_{x_1} x_1 p_{X_1}(x_1) \cdots \sum_{x_n} x_n p_{X_n}(x_n) \\ &= \prod_{j=1}^n \mathbf{E}[X_j] \end{aligned}$$

If discrete random variables X_1, X_2, \dots, X_n are independent, then

$$\begin{aligned}
\mathbf{var}\left(\sum_{j=1}^n X_j\right) &= \mathbf{E}\left[\left(\sum_{j=1}^n X_j - \mathbf{E}\left[\sum_{j=1}^n X_j\right]\right)^2\right] \\
&= \mathbf{E}\left[\left(\sum_{j=1}^n [X_j - \mathbf{E}[X_j]]\right)^2\right] && \text{by the linearity of expectation} \\
&= \mathbf{E}\left[\left(\sum_{j=1}^n \tilde{X}_j\right)^2\right] && \text{for } \tilde{X}_j = X_j - \mathbf{E}[X_j] \\
&= \mathbf{E}\left[\sum_{j=1}^n \sum_{i=1}^n \tilde{X}_j \tilde{X}_i\right] \\
&= \sum_{j=1}^n \sum_{i=1}^n \mathbf{E}\left[\tilde{X}_j \tilde{X}_i\right] && \text{by the linearity of expectation} \\
&= \sum_{j=1}^n \left(\mathbf{var}(X_j) + \sum_{i \neq j} \mathbf{E}\left[\tilde{X}_j \tilde{X}_i\right]\right) && \text{by } \mathbf{var}(\tilde{X}_j) = \mathbf{var}(X_j) \\
&= \sum_{j=1}^n \left(\mathbf{var}(X_j) + \sum_{i \neq j} \mathbf{E}\left[\tilde{X}_j\right] \mathbf{E}\left[\tilde{X}_i\right]\right) && \text{by independence} \\
&= \sum_{j=1}^n \mathbf{var}(X_j) && \text{because } \mathbf{E}\left[\tilde{X}_j\right] = 0. \quad (6.62)
\end{aligned}$$

Example 6.28 (Moments of Binomial Random Variables). Recall that in Example 6.4, we have demonstrated that the binomial random variable discussed parameters n and p can be expressed as the sum of n Bernoulli random variables with parameter p , see (6.6), i.e.,

$$Y = \sum_{j=1}^n X_j.$$

Then as a result the linearity of the expectation (6.25) we have

$$\begin{aligned}
\mathbf{E}[Y] &= \sum_{j=1}^n \mathbf{E}[X_j] && \text{by the linearity of the expectation} \\
&= n \cdot p && \text{because } \mathbf{E}[X_j] = p
\end{aligned}$$

The Bernoulli random variables X_1, \dots, X_n are independent. Thus the sum of their variance is equal to the variance of their sum. Thus

$$\begin{aligned}
\mathbf{var}(Y) &= \sum_{j=1}^n \mathbf{var}(X_j) && \text{by (6.62) because } X_1, \dots, X_n \text{ are independent.} \\
&= n \cdot (1 - p) \cdot p && \text{because } \mathbf{var}(X_j) = p - p^2.
\end{aligned}$$

Let X_1, X_2, \dots independent identically distributed random variables with the common probability mass functions p_X , i.e.,

$$\begin{aligned} p_{X_1^n}(x_1^n) &= \prod_{i=1}^n p_{X_i}(x_i) \quad \forall x_1^n \quad \text{because } X_1, \dots, X_n \text{ are independent,} \\ &= \prod_{i=1}^n p_X(x_i) \quad \forall x_1^n \quad \text{because } X_1, \dots, X_n \text{ are identically distributed with pmf } p_X \end{aligned}$$

The sample mean S_n of the random variables X_1, \dots, X_n is defined as the average of the random variables X_1, \dots, X_n , i.e.,

$$S_n = \frac{X_1 + \dots + X_n}{n} = \frac{1}{n} \sum_{i=1}^n X_i$$

Note that S_n is a random variable as well.

$$\begin{aligned} \mathbf{E}[S_n] &= \mathbf{E}\left[\frac{1}{n} \sum_{i=1}^n X_i\right] \\ &= \frac{1}{n} \sum_{i=1}^n \mathbf{E}[X_i] && \text{by the linearity of the expectation} \\ &= \mathbf{E}[X] && \text{because } \mathbf{E}[X_i] = \mathbf{E}[X] \text{ for all } i \\ \mathbf{var}(S_n) &= \mathbf{var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) \\ &= \frac{1}{n^2} \mathbf{var}\left(\sum_{i=1}^n X_i\right) \\ &= \frac{1}{n^2} \sum_{i=1}^n \mathbf{var}(X_i) && \text{for independent r.v.'s variance of the sum is equal to sum of the variances,} \\ &= \frac{1}{n} \mathbf{var}(X) && \text{because } \mathbf{var}(X_i) = \mathbf{var}(X) \text{ for all } i \end{aligned}$$

The expectation of the sample mean S_n is equal to the expectation of the random variable X , but variance of the sample mean S_n inversely proportional with n . Note that the variance of a random variable is 0 iff its probability mass function is equal to 1 for a particular realization/value. As we consider larger and larger values of n , the sample mean as a random variable becomes more and more similar to the random variable with zero variance. Thus the values of the sample mean S_n is a good estimate of the expected value $\mathbf{E}[X]$. We will discuss this issue more in more detail while discussing laws of large numbers.

Example 6.29 (Estimating Probabilities by Simulation). In many practical situations calculating the probability of an event analytically is either too hard or impractical. If we have access to a computer that can generate independent trials of the random experiment in question than we can use the frequency of the event as an estimate of its probability.

$$\begin{aligned} X_i &= \begin{cases} 1 & \text{if the event } A \text{ occurs in the } i^{\text{th}} \text{ trial.} \\ 0 & \text{if the event } A \text{ does not occur in the } i^{\text{th}} \text{ trial.} \end{cases} \\ p_{X_i}(k) &= \begin{cases} \mathbf{P}(A) & \text{if } k = 1 \\ 1 - \mathbf{P}(A) & \text{if } k = 0 \end{cases} \\ \mathbf{E}[X_i] &= \mathbf{P}(A) && \mathbf{var}(X_i) = \mathbf{P}(A)(1 - \mathbf{P}(A)) \\ S_n &= \frac{X_1 + \dots + X_n}{n} \\ \mathbf{E}[S_n] &= \mathbf{P}(A) && \mathbf{var}(S_n) = \frac{\mathbf{P}(A)(1 - \mathbf{P}(A))}{n} \end{aligned}$$

Example 6.30. Independence is often an assumption in a probabilistic model that is quite accurate for a given range of parameters. Consider for example a city with population n . A fraction of them, say α , are supporters of local soccer team A and the rest are supporters of the rival team B . We want to estimate α . We choose k resident one by one uniformly at random and ask them which local team they support. If the i^{th} resident chosen supports team A then X_i is one otherwise X_i is zero.

$$X_i = \begin{cases} 1 & \text{if } i^{\text{th}} \text{ resident chosen supports team } A, \\ 0 & \text{else} \end{cases}$$

Note that X_i 's are not independent. If the first resident chosen supports team A the second resident, which is chosen after the first one will less likely to be a supporter of team A . However, this dependence is essentially negligible provided that number of residents chosen are much smaller than population, i.e. $k \ll n$. Let us demonstrate this fact by calculating the joint pmf of the random variables X_1^k 's.

Let Y be the total number of team A supporters among k resident chosen, i.e. $Y = \sum_{i=1}^k X_i$. There are $\binom{n}{k}$ ways to choose k residents and $\binom{\alpha n}{y} \binom{(1-\alpha)n}{k-y}$ ways to choose y supporters of team A and $k-y$ supporters of team B . Thus

$$\begin{aligned} p_Y(y) &= \frac{\binom{\alpha n}{y} \binom{(1-\alpha)n}{k-y}}{\binom{n}{k}} \\ &= \frac{(\alpha n)!}{y!(\alpha n - y)!} \cdot \frac{((1-\alpha)n)!}{(k-y)!((1-\alpha)n + y - k)!} \cdot \frac{(n-k)!k!}{n!} \\ &= \binom{k}{y} \frac{\alpha n \cdot (\alpha n - 1) \cdots (\alpha n - y) \cdot (1-\alpha)n \cdot ((1-\alpha)n - 1) \cdots ((1-\alpha)n - (k-y))}{n \cdot (n-1) \cdots (n-k)} \\ &= \binom{k}{y} \alpha^y (1-\alpha)^{k-y} \left(\prod_{j=0}^{y-1} \left(1 - \frac{(1-\alpha)j}{\alpha(n-j)} \right) \right) \left(\prod_{i=0}^{k-y-1} \left(1 + \frac{(1-\alpha)y - \alpha j}{(1-\alpha)(n-y-j)} \right) \right) \end{aligned}$$

Note that there are $\binom{k}{y}$ different ways to choose the indices of y team A supporters among k residents chosen. Thus

$$\begin{aligned} p_{X_1^k}(x_1^k) &= \frac{1}{\binom{k}{y}} p_Y(y) && \text{where } y = \sum_{i=1}^k x_i \\ &= \alpha^y (1-\alpha)^{k-y} \left(\prod_{j=0}^{y-1} \left(1 - \frac{(1-\alpha)j}{\alpha(n-j)} \right) \right) \left(\prod_{i=0}^{k-y-1} \left(1 + \frac{(1-\alpha)y - \alpha j}{(1-\alpha)(n-y-j)} \right) \right) \\ &= \left[\prod_{i=1}^k p_{X_1}(x_i) \right] \cdot \left(\prod_{j=0}^{y-1} \left(1 - \frac{(1-\alpha)j}{\alpha(n-j)} \right) \right) \left(\prod_{i=0}^{k-y-1} \left(1 + \frac{(1-\alpha)y - \alpha j}{(1-\alpha)(n-y-j)} \right) \right) \\ &\approx \prod_{i=1}^k p_{X_1}(x_i) \end{aligned}$$

Along as $k \ll n$ the pmf $p_{X_1^k}$ is very close to that of independent identically distributed random variable with the marginal pmf p_{X_1} .

References

[BT08] Dimitri P. Bertsekas and John N. Tsitsiklis. *Introduction to Probability*. Athena Scientific, Belmont, MA, 2nd edition, 2008.