A Probability Refresher

1 Introduction

The word *probability* evokes (in most people) nebulous concepts related to uncertainty, "randomness", etc.

Probability is also a concept which is hard to characterize formally. The temptation is to define it in terms of frequency of events in repeated experiments, but, as we shall see later, this approach leads to a circular definition: we would end up defining probability in terms of probability.

Instead, as we did with numbers, **we will define probability in terms of axioms**. We will also develop the necessary knowledge to read papers in the field of Information Theory, that often will make explicit use of probabilistic notation.

Section 2 contains the "new and hard" material of this introduction, and rigorously defines probability. You will not be responsible for it in the course.

Section 3 contains a refresher of the elementary probability notions that you should have learned in a prerequisite class, and that will come handy throughout the course. *You should be familiar with the notions and theorems listed in this section.*

Section 4 contains a brief discussion of random variables, and lists some of the most important definitions and theorems, known from elementary probability courses.

Section 5 extends the concepts introduced in the previous section to multiple random variables, and addresses the concept of independence.

Section 6 deals with the topic of expectation, which is really nothing but integration, and introduces the concept of moments.

Section 7 is really an appendix, and deals with simple counting theorems.

2 Axiomatic definition of probability

2.1 Measurable Spaces

Definition: Sample Space . We start or journey towards the definition of probability by introducing a set Ω, called the *sample space* or the *sure event*, which, in this course, is the collection of all possible **outcomes** of an experiment. [1]

[1] Note that *possible* is not a probabilistic concept: an outcome is possible if it can occur, impossible if it cannot occur. For example, if the experiment is the measure of the voltage between two points of a circuit,

■

In general, we will not be concerned with the probability of individual outcomes of an experiment, but with collections of outcomes, called *events*. For instance, when we send a message across a channel, we will be interested in the number of errors in the received signal, rather than in the individual errors. Thus, probability will be defined as a function on sets. If Ω is uncountable, then some of its subsets might be extremely ugly (really, really ugly), and we need suitable collections of sets to work with. These are called σ-algebras.

Definition: Algebra. A collection Σ of subsets of Ω is called an *algebra* on Σ if it has the following three properties:

- $\Omega \in \Sigma$;
- $F \in \Sigma \;\Rightarrow\; F^c \in \Sigma$;
- $F \in \Sigma,\ G \in \Sigma \;\Rightarrow\; F \bigcup G \in \Sigma$.

■

Here F^c is the complement of F in Ω (also denoted as $\overline{F}$), i.e., the set of all elements of Ω that do not belong to F. Note that from the three properties it follows that an algebra is closed under intersection, and therefore it is closed (stable) under finitely many set operations.

Note The term *field* is often used instead of algebra.

Definition Measurable sets If set A belongs to Σ, it is said to be Σ**−measurable**.

■

Definition: σ-algebra. A collection $\mathcal{F}$ of subsets of Ω is a σ-algebraif

- $\mathcal{F}$ is an algebra on Ω;
- if F_n is a countable collection of sets, $n = 1, 2, \ldots$, such that $F_n \in \Omega$ for all n, then $\bigcup_n F_n \in \mathcal{F}$.

■

Ω can be identified with the set of real numbers. Not knowing a priori what the circuit is, we cannot bound the maximum value of the voltage, so we will say that any real number is a possible outcome. However, the result of the experiment will not be a letter of the English alphabet, and no letter is a possible outcome of the experiment.

Thus, a σ-algebra is closed with respect to a *countably many* set operations [2].

What is the intuition behind a σ-algebra? If Ω represents the collection of possible outcomes of an experiment, a subset of Ω is called an **event**. Then, a σ-algebra represents the collection of all possible, *interesting* events from the viewpoint of a given experiment.

Example Let the experiment be a coin toss (where we blow on the coin if it stands up straight!). The set Ω will contain two elements: H and T. The σ-algebra on Ω will contain the following four sets: $\emptyset$, $\{H\}$, $\{T\}$, $\{H, T\}$, where $\emptyset$ is the empty set.

How does one create a σ-algebra? Often one can start from a collection of events, as we did in the above example.

Definition The **σ-algebra generated by a collection C of subsets of Ω** is the smallest σ-algebra on Ω containing the collection C[3].

■

Example Let the experiment be an infinite repetition of coin tosses, which is the simplest example of a source generating infinite sequences of symbols. In Information Theory we love infinite sequences, since we use them to prove many of our main results. Now, there are uncountably many infinite sequences of H and T in Ω. We will define a σ-algebra on Ω by considering the following sets: $\{\omega \mid \omega_i = H\}$ and $\{\omega \mid \omega_i = T\}$ for $i = 1, 2, \ldots$. So, for example, the first of these sets is the collection of sequences that start with H, the 4th is the the collection of sequences that have a T in the 2nd position etc. From these sets we can derive numerous other sets, by applying a finite or countable number of set operation (union, intersection, complement). For instance, the intersection of the 1st and 3rd sets defined above correspond to the collection of sequences that start with HH. By adding the empty set and Ω to the recipe, we get a nice σ-algebra.

Example Often one in the literature encounters the terms **Borel** σ-algebra, and **Borel set**. The Borel σ-algebra on Ω is often denoted as $\mathcal{B}(\Omega)$.
If Ω is a topological space, the Borel σ-algebra on Ω is the σ-algebra generated by the family of open subsets of Ω.
For instance, every subset of $\mathbb{R}$ that we commonly use is a Borel set, One cannot construct non-Borel sets $\in \mathbb{R}$ without the of the of the axiom of choice.

Good News! While the above mechanisms are needed in general, for most applications we can prove theorems for much simpler structures, and the properties carry over to σ-algebras. In general, all we need is a family of sets that is closed under complement and finite intersection (called a π-system).

Example More repeated coin tosses: instead of working with the complex σ-algebra defined above, we will be able to deal with the generating collection (containing $\{\omega \mid \omega_i = H\}$

[2] If Ω has finite cardinality, then there is no difference between an algebra and a σ-algebra. However, if Ω is infinite, the sets in a σ-algebra can be significantly more complex than the sets in an algebra.

[3] The σ-algebra generated by C is the intersection of all the σ-algebras containing C.

and $\{\omega \mid \omega_i = T\}$ for $i = 1, 2, \ldots$) and the finite intersection of its components!

2.2 Probability Spaces

Having defined the sure event and σ-algebras, we now put them together.

Definition Measurable Space A pair $(\Omega, \mathcal{F})$ where Ω is a set and $\mathcal{F}$ is a σ-algebra, is called a **measurable space**. An element of $\mathcal{F}$ (i.e., an event) is called a **F-measurable subset** of Ω.

■

Recall that when dealing with experiments, we are not interested in individual outcomes, but rather in collections of outcomes. It is on such collections that we will define probability. The first step is to consider a non-negative *set function* on an algebra Σ, i.e., a function from the elements of Σ to the non-negative real numbers.

Definition A set function f is **additive** if

- $f(\emptyset) = 0$,
- $f(F \bigcup G) = f(F) + F(G), \ \ \forall F, G | F \bigcap G = \emptyset$.

.

Definition A set function f is **countably additive** if

- it is additive
- if $\{S_n\}$ is a sequence of sets in Σ, and $\bigcup_n S_n \in \Sigma$ (Note: Σ is an algebra, not a σ-algebra!), such that the sets S_n are disjoint, then

$$f(\bigcup_n S_n) = \sum_n f(S_n).$$

Combining a measurable space and a measure, we obtain: **Definition: Measure Space.** A *measure space* is a triple $(\Omega, \mathcal{F}, f)$, where $(\Omega, \mathcal{F})$ is a measurable space and f is a countably additive non-negative set function on $\mathcal{F}$ [4].

■

[4] **Definition** Given a measurable space $(\Omega, \mathcal{F}, f)$, the measure f is **finite** if $f(\Omega) < \infty$.
Definition Given a measurable space $(\Omega, \mathcal{F}, f)$, the measure f is **σ-finite** if there is a sequence $\{S_n\}$of subsets of Ω such that $f(S_n) < \infty$ and $\bigcup_n S_n = \Omega$.
Note Non$-\sigma-$finite measures are a mess! But we do not have to care about them in our course.

And now: **Definition: Probability Measure.** A *probability measure* (finally !) is a $\sigma-$finite measure $\mathbb{P}$ that satisfies

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

■

Putting together a measurable space and a probability measure, we get:

Definition: Probability Space. A *probability space* or *probability triple* is a measure space $(\Omega, \mathcal{F}, \mathbb{P})$where the measure μ is a probability measure.

■

Definition A property that holds on a set S of elements of Ω satisfying $\mathbb{P}(S) = 1$ is said to hold $\mathbb{P}-$**almost everywhere** or **almost everywhere** or $\mathbb{P}-$**almost surely** or **almost surely**. Almost everywhere is abbreviated as **a.e.**, and almost surely is abbreviated as **a.s.**. A synonim often encountered in the literature is **with probability 1**, abbreviated as **w.p. 1**.

2.3 Summary

When modeling an experiment where uncertainty exists, one uses a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

- Ω is a set called the **sample space**.
- Any element $\omega \in \Omega$ is called a **sample point**
- The σ-algebra$\mathcal{F}$ is a family of sets called **events**. Therefore, and event is an $\mathcal{F}-$measurable set.
- The probability $\mathbb{P}$ is a probability measure.

For a given experiment, we can think of a map between Ω and the set of possible outcomes, where each $\omega \in \Omega$ correspond to only one outcome. In general, this mapping could be a many-to-one. For instance, in a coin flip, Ω could be the set of all configurations of the coin when it is released (= position, speed and angular speed), and of all the molecules that can influence the outcome. Clearly Ω is a rather large set, while the set of possible outcomes is very small. Many ω map to head, and the remaining to tail. Or, to make things simpler, we can just think of Ω as the collection $\{H, T\}$.

3 Elementary Probability Refresher

We can now put behind us the complexity of the notions outlined in the previous section, and recall few notions from elementary probability. These results are stated without proof.

3.1 Basic properties of probability

Let A and B be sets. Then the following properties hold.

- $0 \leq \mathbb{P}(A) \leq 1$
- $\mathbb{P}(A) = 1 - \mathbb{P}(A^c)$, where $A^c)$ is the complement of A;
- $\mathbb{P}(\emptyset) = 0$;
- $A \subseteq B$ implies $\mathbb{P}(A) \leq \mathbb{P}(B)$;
- If $A_1, \ldots, A_n$, with n finite, are disjoint, then $\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathbb{P}(A_i)$.
- If $A_1, A_2, \ldots$ is a countable sequence of disjoint sets, $\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

Theorem Inclusion/Exclusion Principle. For any pair of events A and B,

$$\mathbb{P}(A \bigcup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \bigcap B).$$

■

Theorem Union of Events Bound. For any pair of events A and B,

$$\mathbb{P}(A \bigcup B) \leq \mathbb{P}(A) + \mathbb{P}(B).$$

Note that the union of event bound is a corollary of the Inclusion/Exclusion principle.

■

Definition: Conditional Probability.. Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, let A and B be two events in $\mathcal{F}$. Assume that we know that B occurs. The probability that A occurs given that it is known that B occurs is called the *conditional probability of A given B* and is defined by

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \bigcap B)}{\mathbb{P}(B)}.$$

■

Example Consider a fair die, the conditional probability that an outcome X is odd given that it is less than four is $\mathbb{P}(\{1,3\})/\mathbb{P}(\{1,2,3\}) = 2/3$.

Note Recall that two events are said to be independent if the probability of their intersection is the product of their individual probabilities. Under the assumption that $\mathbb{P}(B) > 0$, it follows immediately that $\mathbb{P}(A \mid B) = \mathbb{P}(A)$ if and only if A and B are independent.

From the definition of conditional probability, we immediately obtain:

Theorem - Multiplication law. Let A and B be events with $\mathbb{P}(B) > 0$. Then

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

■

The following is a very useful theorem, which is a consequence of the multiplication law.

Bayes Theorem. Given two events A and B,

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

■

Total Probability Theorem
Let $A_1, A_2, \ldots$ be a sequence (possibly infinite) of *disjoint* events (i.e., for each i, j, with $i \neq j$, $A_i \bigcap B_j = \phi$), such that $\bigcup_{i=1}^{\infty} A_i = \Omega$. Then, for every measurable event B,

$$\mathbb{P}(B) = \mathbb{P}(\bigcup_{i=1}^{\infty}(B \bigcap A_i) = \sum_{i=1}^{\infty} \mathbb{P}(B \mid A_i)P(A_i).$$

Definition: Independence. Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, two events A and B $\in \mathcal{F}$ are *independent* if

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

■

Thus, two events are called independent if the probability that they both occur is the product of the probabilities that each of them occurs. The definition can be extended to N events, but the extension *is not trivial !*

Definition Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, N events $A_i \in \mathcal{F}$, $i = 1, \ldots, N$ are *independent* if for every $n \geq 2$, every collection of different indices $i_1, \ldots, i_n$ with $1 \leq i_j \leq N$,

$$\mathbb{P}\left(\bigcap_{j=1}^{n} A_{i_j}\right) = \prod_{j=1}^{n} \mathbb{P}\left(A_{i_j}\right).$$

■

Definition A **fair coin** is a coin for which $\mathbb{P}(Head) = \mathbb{P}(Tail) = 1/2$.

Definition A **fair die** is a die for which $\mathbb{P}(X = i) = 1/6$ for $i = 1, \ldots, n$.

Example Consider tossing a fair die, call O the outcome. Let A be the event $\{O$ is odd$\}$ and B be the event $\{O$ is less than 3$\}$. As the die is fair, $\mathbb{P}(A) = 1/2$ and $\mathbb{P}(B) = 1/3$. Note that $A \bigcap B = \{X = 1\}$, thus $\mathbb{P}(A \bigcap B) = 1/6 = \mathbb{P}(A)\mathbb{P}(B)$. Then the events A and B are independent.

Some Properties and Non-properties of independence of events

- Independence is symmetric: the statements "A and B are independent", "A is independent of B", and "B is independent of A" are equivalent.
- A strange case. An event A having zero probability is
  - independent of every other event;
  - independent of the sure event;
  - independent of itself.
- Independence does not mean mutual exclusion! If A and B are events with *non-zero probability*, and are mutually exclusive, i.e., if A occurs, then B does not occur and vice versa, then A and B are **not** independent. In fact if $A \bigcap B = \emptyset$, then $\mathbb{P}(A \bigcap B) = 0$; but, by assumption, $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$, thus $\mathbb{P}(A)\mathbb{P}(B) > 0$. For example, when tossing a die, the events {outcome is odd} and {outcome is even} are not independent.
- Do not think of independence (or dependence) in terms of causality, as it is misleading. Think in terms of probability.

4 Random Variables

Definition A function $h : S \to \mathbb{R}$ is $\Sigma-$**measurable function** if its inverse maps Borel sets of the Real line into elements of the algebra Σ.

An important case is when Σ is itself the Borel σ-algebra.

Lemma Sums and products of measurable functions are measurable functions.

Definition If $(\Omega, \mathcal{F}, \mathbb{P})$is a probability space, we call *random variable* any $\mathcal{F}-$measurable function.

■

Example Let Ω describe the tossing of two dice. The sum of the values is a random variable. So is the difference, the maximum and the minimum.
ATTENTION the maximum and minimum of an infinite collection of random variables need not be measurable, and therefore need not be a random variable. The infimum, liminf and limsup are always random variables, and so is the lim, if it exists.

Example Consider again the example of infinite coin tosses. We define $\mathcal{F}$ as the σ-algebra generated by the sets $\{\omega : \omega_n = W\}$ for $W \in \{H, T\}$ and all $n \in \mathbb{N}$. Then, one can easily show that the following are random variables:

- $X_n(\omega) = 1$ if $\omega_n = H; = 0$ otherwise .
- $Y_n(\omega) = \sum_{i=1}^{n} 1(\omega_i = H)$, i.e., the number of Heads in the first n coin tosses.
- $Z_n(\omega) = 1$ if $Y_n(\omega)$ is odd, $= 0$ otherwise.

4.1 Distributions and Laws

Let $(\Omega, \mathcal{F}, \mathbb{P})$be a probability space carrying a random variable X. Then the inverse mapping X^{-1} is a correspondence between Borel sets B and sets $S \in \mathcal{F}$. The probability measure $\mathbb{P}$ assigns a value between 0 and 1 to each set S.

Definition We define the **Law** $\mathcal{L}_X$ of X as

$$\mathcal{L}_X = \mathbb{P} \circ X^{-1}.$$

Thus, the law of X assigns to each Borel set the probability under $\mathbb{P}$ of its preimage. The law of Z_n assigns probability to the events: $\emptyset$, $\{0\}$, $\{1\}$ and $\{0, 1\}$. The probability of $\emptyset$ is trivially equal to 0, and the probability of $\{0, 1\}$ is trivially 1. The law $\mathcal{L}_{Z_n}$ assigns to the event $\{0\}$ the probability of its preimage, which is the set of all sequences of coin tosses with even number of heads in the first n trials, and to the event $\{1\}$ the probability of the set of all sequences having odd number of heads in the first n trials. The "magic" of probability is now becoming increasingly clear.

GOOD NEWS ! This is the end of the "new material". What comes next is just review material from elementary probability !

Definition: Cumulative Distribution Function. The function $F_X(x) = \mathcal{L}_X\{(-\infty, x]\}$ is called the *(cumulative) distribution function* of X.

■

Thus, the distribution function of X evaluated at x is the probability that X is less than **or equal to** x, which is the measure under $\mathbb{P}$ of the preimage of the set $(-\infty, x]$.

Example Consider tossing a die, and let X be the number on the upper face. The distribution function of X is equal to 0 for every $x < 1$, is equal to the probability of the preimage of $\{X = 1\}$ for $x \in [1, 2)$, to the probability of the preimage of $\{X \in \{1, 2\}\}$ for $x \in [2, 3)$ etc. etc. If the die is fair, then the distribution function of X is equal to zero for $x < 1$, to 1/6 for $x \in [1, 2)$, to 2/6 for $x \in [2, 3)$ to 1/2 for $x \in [3, 4)$ to 2/3 for $x \in [4, 5)$ to 5/6 for $x \in [5, 6)$ to 1 for $x \geq 6$.

Properties of the distribution function

- $F_X \in [0,1]$;
- $F_X(x)$ is monotonically non-decreasing in x, i.e., for all pairs x_1, x_2 with $x_1 < x_2$ it satisfies $F_X(x_1) \leq F_X(x_2)$;
- $F_X(x)$ is right continuous (continuous from the right), i.e., $\lim_{\epsilon \to 0+} F(x+\epsilon) = F(x)$;
- $\lim_{x \to -\infty} F_X(x) = 0$;
- $\lim_{x \to +\infty} F_X(x) = 1$.

Definition: Probability Mass Function. Let a random variable X have a law that puts probability one on a finite or countable subset of the real line $\mathcal{X} = \{x_1, x_2, \ldots\}$ (for instance, the integers) and zero probability on the rest of the real line. This random variable is called **discrete**. With probability one X will take a value that lies in $\mathcal{X}$. The function $P_X(\cdot) = P(X = x_i)$ for $i = 1, \ldots$ is called the *probability mass function* of X.

Definition: Probability Density Function. Let F_X be absolutely continous with respect to the Lebesgue measure on the real line[5]. Then F_X is differentiable everywhere. Let $f_X(\cdot)$ be its derivative. The function $f_X(\cdot)$ is called the probability density function of X, or, more simply, the *density* of X.

■

4.2 Independent Random Variables

Definition Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, on which two random variables X and Y are defined, with laws $\mathcal{L}_X$ and $\mathcal{L}_Y$. X and Y are **independent** if, for every pair of Borel sets A and B, $\mathbb{P}(X \in A; Y \in B) = \mathcal{L}_X(A)\mathcal{L}_Y(B)$. That is, the probability that X takes value in A and Y takes value in B is equal to the product of the probability that X takes value in A and of the probability that Y takes value in B.

■

Note Good news ! We can restrict the attention to simple sets ! The following theorem simplifies life enormously ! (the proof requires some additional machinery, and therefore is omitted.)

Theorem *Two random variables X and Y defined on $(\Omega, \mathcal{F}, \mathbb{P})$, having distribution functions $F_X(\cdot)$ and $F_Y(\cdot)$ are independent if and only if, for every x and y,*

$$\mathbb{P}(X \leq x; Y \leq y) = \mathbb{P}(X \leq x)\mathbb{P}(Y \leq y).$$

■

[5]This means that there are no sets of Lebesgue measure zero that have non-zero probability under the law of X.

Thus, we can concentrate the attention on the probabilities of very simple sets ! We can also define independence of n random variables:

Theorem *N random variables $X_1, \ldots, X_N$ defined on $(\Omega, \mathcal{F}, \mathbb{P})$ are independent if and only if, for every $n \leq N$, every collection of different indices $i_1, \ldots, i_n$, and every collection of real numbers $x_1, \ldots, x_n$,*

$$\mathbb{P}(X_{i_1} \leq x_1; \ldots; X_{i_n} \leq x_n) = \mathbb{P}(X_{i_1} \leq x_1) \ldots \mathbb{P}(X_{i_n} \leq x_n).$$

■

Note The theorem can be restated in a "recursive" fashion as follows: N random variables $X_1, \ldots, X_N$ defined on $(\Omega, \mathcal{F}, \mathbb{P})$ are independent if and only if the following two conditions hold:

- for every collection of N real numbers $x_1, \ldots, x_n$, $\mathbb{P}(X_1 \leq x_1; \ldots; X_N \leq x_N) = \prod_{i=1}^{N} \mathbb{P}(X_i \leq x_i)$,
- (if $N > 2$) every subcollection of $\{X_1, \ldots, X_N\}$ containing $N - 1$ different X_i's, is composed of independent random variables.

Definition The random variables $X_1, X_2, \ldots$ (forming a countable sequence) are independent if any finite size subset of the sequence is composed of independent random variables.

■

Definition: Independent and Identically Distributed Random Variables. Independent random varibles $X_1, \ldots, X_N$ having the same law are said to be *Independent and Identically Distributed (iid).*

■

5 Multiple random variables

In the course we will often deal with groups of random variables, which we will assume to be defined on a common probability space. We extend the definitions of the previous section to a pair of random variables. The extension to a finite number of random variables is trivial.

Definition: Joint Distribution. Let X and Y be two real-valued random variables. Their joint (cumulative) distribution (function) $F_{X,Y}(x, y)$ is defined as

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

■

Definition: Joint Density. The joint density of a pair of variables X and Y with joint distribution $F_{X,Y}(x,y)$Again, we must require that F be absolutely continuous with respect to the Lebesgue measure. is the function

$$f_{X,Y}(x,y) = \frac{d^2}{d\tilde{x}\, d\tilde{y}} F_{X,Y}(\tilde{x},\tilde{y})\Big|_{x,y},$$

(where $\tilde{x}$ and $\tilde{y}$ are differentiation dummy variables).

■

Definition: Marginal. Consider two random variables X and Y, with joint density $f_{X,Y}(x,y)$. The *marginal* distribution of X is

$$f_X(x) = \int_{\mathbb{R}} f_{X,Y}(x,y)\, dy$$

and similarly we define the marginal of Y, $f_Y(y)$.

■

Note Knowing both marginals is **NOT** equivalent to knowing the joint distribution.

Definition: Conditional Density. Given two random variables X and Y, with joint density $f_{X,Y}(x,y)$, the *conditional density of X given that $Y = y$* is

$$f_X(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}.$$

Thus, the conditional density of X given Y is a random function (note the difference from above, here we did not specify $Y = y$!) given by $f_X(x \mid Y) = f_{X,Y}(x,Y)/f_Y(Y)$.

Extension of Theorems

We extend the following theorems to densities and PMF's. In particular, we use the density notation.

- **Bayes Theorem**
$$f_X(x \mid y) = f_Y(y \mid x)\frac{f_X(x)}{f_Y(y)}.$$
- **Total Probability Theorem**
$$f_X(x) = \int f_X(x \mid y) f_Y(y)\, dy.$$

Definition: Independence. Two random variables X and Y defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$are independent if

$$f_{X,Y}(x,y) = f_X(x) f_Y(y),$$

i.e., if the joint density is the product of the marginals.

■

Note If the joint density is the product of the marginals, and $\mathcal{R}$ is the Borel σ-algebra over the reals, then for every $A \in \mathcal{R}$, $B \in \mathcal{R}$,

$$P(X \in A, Y \in B) = P(X \in A)\ P(Y \in B)$$

6 Expectation

Continuous random variable

Let X be a random variable with density $f_X(\cdot)$. We define the **expectation** or **expected value** (or **mean**) of X as

$$E(X) = \int_{\mathbb{R}} x\ f_X(x) dx$$

if it exists, where the integral is meant to be a Lebesgue integral (actually, Lebesgue/Stjelties integral).

■

Note The Riemann Integral will not cut it here !

Note The integral MUST NEVER be interpreted as a principal value integral !!!!!

Discrete Random Variable

Let X be a random variable with probability mass function $P_X(\cdot)$. We define the **expectation** of X as

$$E(X) = \sum_{i=1}^{\infty} x_i\ P_X(x_i),$$

if $\sum_{i=1}^{\infty} |x_i|\ P_X(x_i)$ exists, and is undefined otherwise. Note that summation is just a particular type of integration.

Definition: Expectation of a function of a random variable. Let $g(\cdot)$ be a function, and X a random variable with law $\mathcal{L}_X$. Then $g(X)$ is a random variable too. The expectation

of $g(\cdot)$ with respect to $\mathcal{L}_X$ is the expectation of $g(X)$, can be written in one of the following ways

$$E_{\mathcal{L}_X}(g(X)),\ E_{\mathcal{L}_X}(g(\cdot)),\ E_{F_X}(g(X)),\ E_{F_X}(g(\cdot)),\ E_{f_X}(g(X))E_{f_X}(g(\cdot)),\ E_X(g(X)),\ E_X(g(\cdot))$$

and is defined as

$$E_X(g(\cdot)) = \int_{\mathbb{R}} g(x) f_X(x) dx$$

if X has density $f_X(\cdot)$[6], and as

$$E_X(g(\cdot)) = \sum_{i=1}^{\infty} g(x_i) P_X(x_i),$$

if X has probability mass function $P_X(\cdot)$.

■

Definition: Moments. The kth moment of a random variable X with density $f_X(\cdot)$ is defined as

$$E(X^k) = \int_{\mathbb{R}} x^k f_X(x) dx,$$

if the integral exists, and is undefined otherwise.
The kth moment of a random variable X with probability mass function $P(X)$ is defined as

$$E(X^k) = \sum_{i=1,\ldots} x_i^k P(x_i),$$

if the sum exists, i.e., if $\sum_{i=1,\ldots} \left| x_i^k \right| P(x_i)$ is finite, and is undefined otherwise.

■

Definition: Central Moments. The kth central moment of a random variable X with density $f_X(\cdot)$ and having expectation $E(X)$ is defined as

$$E(X^k) = \int_{\mathbb{R}} [x - E(x)]^k f_X(x) dx,$$

if the integral exists, and is undefined otherwise.
The kth moment of a random variable X with probability mass function $P(X)$ is defined as

$$E(X^k) = \sum_{i=1,\ldots} [x - E(x)]^k P(x_i),$$

if the sum exists, and is undefined otherwise.

■

Example The 2nd central moment of a random variable is called **variance**, the 3rd central moment is called **skewness**.

[6]To be exact, if $f_X(\cdot)$ is the density of X with respect to the Lebesgue measure, we must require that $h(\cdot)$ be a Lebesgue-measurable function.

7 Some Counting Theorems

Let's have an urn, containing n balls each with a different symbol. **Sampling with replacement** consists of repeatedly extracting a ball from the urn, noting the symbol, putting the ball back in the urn, and shaking the urn. Let $\mathcal{X} = \chi_1, \ldots, \chi_n$ be the set of n different symbols written on the balls, and let χ be the outcome of a sampling operation. We will model sampling with replacement using the following assumptions:

- $\mathbb{P}(\chi = \chi_i) = 1/n$,
- sampling operations are independent.

As a consequence, all the possible outcomes of k subsequent sampling with replacement operations have the same probability.

Lemma The number of different ordered outcomes of k sampling with replacement operations from an urn containing n elements is n^k.

Sampling without replacement consist of extracting a ball from an urn, noting the symbol, and setting the ball aside. We will assume that the probability distribution modeling the selection of a ball is uniform. We will also assume that all the possible sequences of balls resulting from k subsequent sampling without replacement operations are equiprobable.

Lemma The number of different **ordered** outcomes of k sampling without replacement operations from an urn containing n items is equal to $n * (n-1) * \ldots * (n-k+1)$, i.e. $n!/(n-k)!$.

Lemma The number of different ordering of n elements is $n!$.

Lemma The number of different **unordered** outcomes of k sampling without replacement operations from an urn containing n items is

$$C(n,k) = \left(\begin{array}{c} n \\ k \end{array} \right) = \frac{n!}{k!(n-k)!}.$$