11

PROBABILITY THEORY

11.1 σ algebra

11.1.1 σ algebra concepts

Definition 11.1.1 (σ algebra). *Given a set Ω, a σ-**field**, or σ-**algebra** is a collection $\mathcal{F}$ of subsets of Ω, with the following properties:*

1. $\emptyset \in \mathcal{F}$
2. *if $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$*
3. *(countable union)if $A \in \mathcal{F}$, then $\cup_{i=0}^{\infty} A_i \in \mathcal{F}$*

Example 11.1.1.

1. The trivial σ-field $\mathcal{F} = \{\emptyset, \Omega\}$
2. The collection $\mathcal{F} = \{\emptyset, A, A^c, \Omega\}$, where A is a fixed subset of Ω
3. The set of all the subsets of finite set Ω.
4. For a finite sample space Ω, the power set of Ω is the largest σ field, $\{\emptyset, \Omega\}$ is the smallest σ field.

Remark 11.1.1. The pair $(X, \mathcal{F})$ is called **measurable space**, the members $e \in \mathcal{F}$ are called **measurable sets** or $\Sigma-$measurable sets.

Lemma 11.1.1 (intersection theorem). *[1] If $\{\mathcal{F}_\alpha\}_{\alpha \in T}$ is a collection of σ fields on Ω, then $\cap_{\alpha \in T} \mathcal{F}_\alpha$ is σ field on Ω*

Proof. We consider the special case $T = \{1, 2\}$. Let $A = \mathcal{F}_1 \cap \mathcal{F}_2$. It is easy to see $\emptyset \in A$; since $A \in \mathcal{F}_1 \cap \mathcal{F}_2$, then $A \in \mathcal{F}_1, A \in \mathcal{F}_2$, then $A^c \in \mathcal{F}_1, A^c \in \mathcal{F}_2$, then $A^c \in \mathcal{F}_1 \cap \mathcal{F}_2$; Similarly, we can prove the union property. □

11.1.2 Generation of sigma algebra

Lemma 11.1.2 (Existence of smallest σ field, σ algebra generation). *If $\mathcal{A}$ is a collection of subsets of Ω, then there exist a unique smallest σ field on Ω, containing $\mathcal{A}$, which is contained by all the σ fields that contains $\mathcal{A}$. We denote this by $\mathcal{F}(\mathcal{A})$, and called the σ field generated by $\mathcal{A}$.*

Proof. Consider $\mathcal{B}$ as the set of all σ fields that contains $\mathcal{A}$. The intersections of all these sets will lead to $\mathcal{F}(\mathcal{A})$ due to Lemma 11.1.1. □

Definition 11.1.2 (sigma algebra generated by an event). *Let A be a subset of a set Ω. The sigma algebra generated by A, denoted by $\sigma(A)$, is a set given by*

$$\sigma(A) = \{\emptyset, \Omega, A, A^c\}.$$

Remark 11.1.2 (sigma algebra generated by random variable and stochastic process). The generation of sigma algebra by random variables and stochastic processes are discussed in Definition 11.3.3Definition 18.6.3.

Corollary 11.1.0.1 (Properties of generated σ algebra). *[1] If $\mathcal{A}, \mathcal{A}_1$ and $\mathcal{A}_2$ are subsets of 2^Ω, then we have*

- *If $\mathcal{A}_1 \subset \mathcal{A}_2$, then $\mathcal{F}(\mathcal{A}_1) \subset \mathcal{F}(\mathcal{A}_2)$*
- *If $\mathcal{A}$ is a σ field, then $\mathcal{F}(\mathcal{A}) = \mathcal{A}$*
- *If $\mathcal{F}(\mathcal{F}(\mathcal{A})) = \mathcal{F}(\mathcal{A})$*

11.1.3 Partition of sample space

Definition 11.1.3 (partition of sample space). *A collection of subsets of Ω, $\{\mathcal{A}_i\}_{i\in I}$ (I can have size of uncountable infinite)is called a partition of Ω if*

$$\mathcal{A}_i \cap \mathcal{A}_j = \emptyset, \text{ if } i \neq j$$

and

$$\cup A_i = \Omega.$$

Lemma 11.1.3. *[1] If $\mathcal{P} = \{A_i\}_{i=1}^\infty$ is a countable partition of Ω, then the σ field generated from $\mathcal{P}$, $\mathcal{F}(\mathcal{P})$, consists of all sets of the form $\cup_{n\in M} A_n$ where M ranges over all subsets of $\mathbb{N}$.*

Lemma 11.1.4. *Let $\mathcal{P}_1, \mathcal{P}_2$ be the partitions of the same set Ω. If $\mathcal{P}_2$ is obtained by subdividing sets in $\mathcal{P}_1$ (i.e. $\mathcal{P}_2$ is finer), then we have*

$$\mathcal{F}(\mathcal{P}_1) \subseteq \mathcal{F}(\mathcal{P}_2) \Leftrightarrow \mathcal{P}_1 \subseteq \mathcal{P}_2$$

11.1.4 Filtration & information

Definition 11.1.4 (filtration). *Let $(\Omega, \mathcal{F})$ denote a measurable space.*

- *A **continuous filtration** is defined as: A family of σ algebras $\{\mathcal{F}_t | t >= 0\}$ where*

$$\mathcal{F}_s \subseteq \mathcal{F}_t \subseteq \mathcal{F}, 0 \leq s \leq t$$

- *A **discrete filtration** on $(\Omega, \mathcal{F})$ is an increasing sequence of σ fields $\{\mathcal{F}_n\}$ such that:*

$$\mathcal{F}_1 \subseteq \mathcal{F}_2 ... \subseteq \mathcal{F}$$

Note that as the time progresses, the finer the σ algebra will be. We call $\mathcal{F}_n$ the history up to time n.

Remark 11.1.3. Note that usually not all the subsets of X can be defined a measure with above properties. For example, all irrational numbers in the real line, the root to polynomial equation, are not measurable sets.[2]

Remark 11.1.4 (filtration and information).

- Let $\mathcal{F}_1, \mathcal{F}_2$ be two σ field on Ω, then $\mathcal{F}_1 \subseteq \mathcal{F}_2$ mean $\mathcal{F}_2$ contains more information than $\mathcal{F}_1$; For any A is measurable with respect to $\mathcal{F}_1$ then A is measurable with respect to $\mathcal{F}_2$. That is, if $A \subset \mathcal{F}_1$ then $A \subset \mathcal{F}_2$.

Example 11.1.2. For example, in a die toss example, $\mathcal{F}_1$ is generated by the events of odd number or even number, while $\mathcal{F}_2$ is generated by the event of all possible outcomes. Then, we have $\mathcal{F}_1 \subset \mathcal{F}_2$,i.e., knowing the probability measure on $\mathcal{F}_2$ will enable us to calculate the probability measure on $\mathcal{F}_1$.[1]

Now consider a series of experiment: Let Ω denote the set of all outcomes resulting from tossing a coin three times, the $\Omega = \{(H,H,H),(T,H,H)...,(T,T,T)\}$. Let $\mathcal{F}_i$ denote the events that have been determined by the end of the i toss. Then $\mathcal{F}_1 = \mathcal{F}(\{(H,\cdot,\cdot),(T,\cdot,\cdot)\})$, where $\cdot$ represent it will range over H, T,i.e., $\mathcal{F}_1$ is generated from a partition of 2. Since we have more information later, we have

$$\mathcal{F}_1 \subset \mathcal{F}_2 \subset \mathcal{F}_3$$

Note that if $\mathcal{F}_2$ represents events determined by the i toss instead of tosses upto i, the above will not hold.

11.1.5 Borel σ algebra

Definition 11.1.5. *[3]*

- *A **Borel set** is a set in a topological space that can be formed from open sets (or from closed sets) through the operations of countable union, countable intersection, and relative complement.*
- *For a topological space X, the collection of all Borel sets on X forms a σ-algebra, known as the **Borel σ-algebra**. The Borel σ algebra on X is the smallest σ-algebra generated by open sets.*

Remark 11.1.5. Note that the elements like low-dimensional manifold $S \subset \mathbb{R}^m, m < n$ in $\mathbb{R}^n$ will not be in the $\mathcal{B}(\mathbb{R}^n)$, i.e., they cannot be obtained from open set operation defined above.

Note 11.1.1 (open interval close interval conversion). Using countable union and intersection properties, we can convert between open interval and close intervals, for example

- $(a, b) = \cup_{n=1}^{\infty}[a + 1/n, b - 1/n]$
- $[a, b] = \cap_{n=1}^{\infty}(a - 1/n, b]$
- $(a, \infty) = \cup_{n=1}^{\infty}[a, a + n]$
- singleton: $\{a\} = [a, a]$

11.1.6 Measurable set and measurable space

A **measure** on a set is a systematic way to assign a number of each suitable subset of set, as a generalization of the concepts of length, area, and volume.

Definition 11.1.6 (measure). *Given a set X with its σ field Σ, a function $\mu : \Sigma \to \mathbb{R}$ is called a **measure** if it satisfies:*

- *Non-negativity: For all $E \in \Sigma$, $\mu(E) >= 0$*
- *$\mu(\emptyset) = 0$*
- *Countable additivity: For all countable collections $\{E_i\}$ of pairwise disjoint sets in Σ:*

$$\mu(\cup_{i=1} E_i) = \sum_{i=1} \mu(E_i)$$

*The pair (X, Σ) is called **measurable space**, the members $e \in \Sigma$ are called **measurable sets** or $\Sigma-$measurable sets. A triple (X, Σ, μ) is called measure space.*

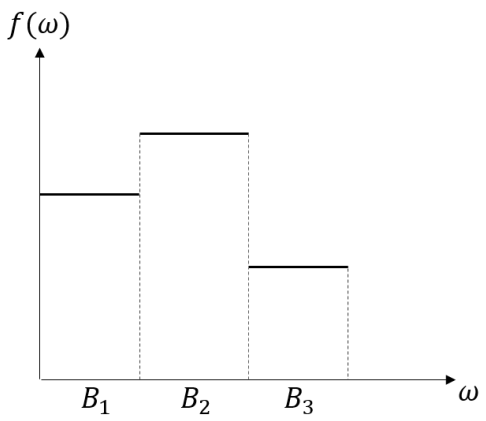
Example 11.1.3 (probability measure). A **probability measure** is a measure satisfying above three properties and has one additional requirement of total measure one $\mu(X) = 1$.

Definition 11.1.7 (measurable function, Borel measurable function). *Let $(\Omega, \mathcal{F})$ be a measurable space. A function $f : \Omega \to \mathbb{R}$ is said to be $\mathcal{F}$-measurable, or Borel measurable, if $f^{-1}(B) \in \mathcal{F}, B \in \mathcal{B}(\mathbb{R})$.*

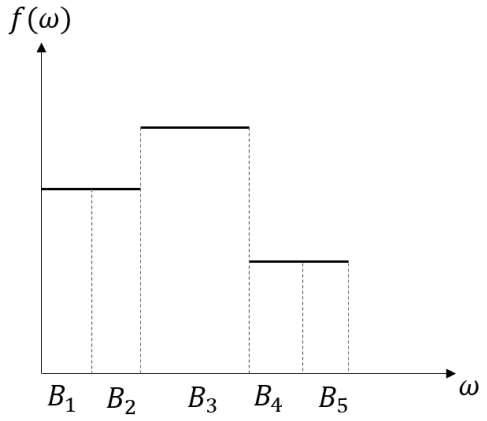
Example 11.1.4 (measurable function with coarse sigma field). Let $\mathcal{F}$ generated by a finite partitions $B_1, B_2, ..., B_m$ of Ω; let function $f : \Omega \to \mathbb{R}$ be **$\mathcal{F}$-measurable**. Then f take constant value on each element of $B_i, 1 \leq i \leq m$ [Figure 11.1.1].

Suppose f can take different values, say a_1, a_2, then the inverse image of the interval $[a_1, 0.5(a_1 + a_2)]$ is not a subset of $\mathcal{F}$ (note that $\mathcal{F}$ can only contain $\emptyset$ plus subsets due to unions of partition subset. See previous sections on partition of sample space), which contradicts the fact of f is measurable.

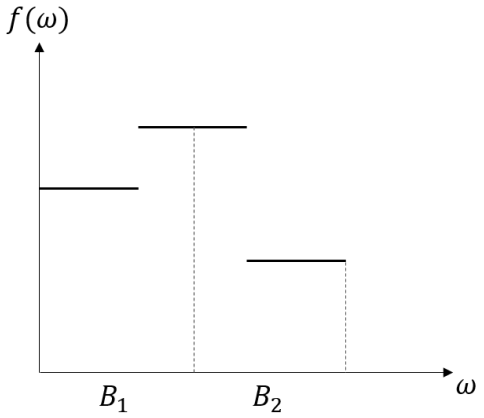
Therefore measurability usually limits the 'variation' of a function defined on a set.



(a) A mesurable function measurable with respect to the σ algebra generated by partition B_1, B_2, B_3.



(b) A mesurable function measurable with respect to the σ algebra generated by partition B_1, B_2, B_3, B_4, B_5.



(c) A function not measurable with respect to the σ algebra generated by partition B_1, B_2.

Figure 11.1.1: An illustration of measurable functions.

Note 11.1.2 (measurable functions vs. ordinary functions)**.**

- Ordinary functions from set A to set B simply establish a relationship between elements in A and elements in B. A measurable function from set A to set B also establish a relationship between elements in A and elements in B, however, under the constraint of measurability of σ fields.
- The level of coarseness constrain the number of values a measurable function can take. For a trivial $\mathcal{F}$, its measurable function can only take one value.
- For random variables, they are required to be measurable functions.

11.2 Probability space

11.2.1 Event, sample point and sample space

Consider a experiment that gives random outcomes. The results of the experiment or observations are called **events**. For example, the result of a measurement will be called an event. We shall distinguish between *compound*(or decomposable) and *simple*(or indecomposable) *events*. For example, saying that a throw with two dice resulted in "sum six" amounts to saying that it resulted in (1,5) or (2,4) ..., which can be decomposes to five simple events. The simple events will be called sample points. Every dis-decomposable result of the experiment is represented by one and only one, sample point. The aggregate of *all* sample points is called **sample space**.

More formally, we have following definition.

Definition 11.2.1 (event,sample point and sample space). *Consider a random experiment. The collection of all outcomes is the sample space Ω. Given a sample space Ω with its σ field $\mathcal{F}$, an event is simply an element in $\mathcal{F}$.*

11.2.2 Probability space

Definition 11.2.2 (probability space). *A probabilistic model is defined by a triple $(\Omega, \mathcal{F}, P)$, called a **probability space**, where*

1. *Ω is the sample space, the set of possible outcomes of the experiment.*
2. *$\mathcal{F}$ is a σ-field, a collection of subsets of Ω, containing Ω itself and the empty set $\emptyset$, and closed under the formation of complements, countable unions, and countable intersections.*
3. *P is a **probability measure** defined on σ-field $\mathcal{F}$, and has the property of:*
   - *$P(A) \geq 0, \forall A \in \mathcal{F}$*
   - *if $A_1, A_2, ... \in \mathcal{F}$ are **disjoints** subsets of Ω, we have **countable additivity** as: $P(A_1 \cup A_2 \cup A_3 ...) = P(A_1) + P(A_2) + P(A_3) + ...$*
   - *$P(\Omega) = 1$.*

Remark 11.2.1 (interpretation).

- Note that the σ-algebra is the collections of *measurable sets*. These are the subsets $A \subseteq \Omega$ where $P(A)$ is defined. In general, σ-field might not contain *all* subsets of Ω. For example, let Ω be a interval on the real line, then the set of all rational number in the interval is not in σ-field.

- **Note that we cannot extend to *uncountable unions*;** in this case, $\mathbb{F}$ would contain every subset A, since every subset can be written as $A = \cup_{x \in A}\{x\}$ and since the singleton sets $\{x\}$ are all in $\mathbb{F}$.

Definition 11.2.3 (discrete probability space). *A **discrete probability space** is a triplet $(\Omega, \mathbb{F}, \mathbb{P})$ such that*

1. *the sample space is finite or countable.*
2. *the σ-field is the set of all subsets of Ω.*
3. *the probability measure (a function) assigns a number in the set [0,1] to every pairwise disjoint subset of $A \subseteq \Omega$, given by*

$$\mathbb{P}(A) = \sum_{\omega \in A} \mathbb{P}(\{\omega\}),$$

and

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

Example 11.2.1 (Infinite coin toss process (infinite Bernoulli experiments)). [4, p. 4]

- Consider the probability space for tossing a coin infinitely many time. We can define the sample space as $\Omega_\infty =$ the set of infinite sequences of Hs and Ts. A generic element of Ω_∞ will be denoted as $\omega = \omega_1\omega_2$...., where ω_n indicates the result of the nth coin toss $\omega_n = H$ *or* T.
- Example subsets in Ω are
  - A_H: the set of all sequences beginning with H. $A_H = \{\omega : \omega_1 = H\}$.
  - A_T: the set of all sequences beginning with H. $A_T = \{\omega : \omega_1 = H\}$.
  - A_{HT}: the set of all sequences beginning with HT. $A_{HT} = \{\omega : \omega_1 = H, \omega_1 = T\}$.
  - A_{TH}: the set of all sequences beginning with TH. $A_{TH} = \{\omega : \omega_1 = T, \omega_2 = H\}$.
- Possible σ algebra includes:
  - $\mathcal{F}_0 = \{0, \Omega_\infty\}$.
  - $\mathcal{F}_1 = \{0, \Omega_\infty, A_H, A_T\}$.
  -

$$\mathcal{F}_2 = \begin{array}{l} 0, \Omega_\infty, A_H, A_T, A_{HH}, A_{HT}, A_{TH}, A_{TT}, A_{HH}^C, A_{HT}^C, A_{TH}^C, A_{TT}^C, \\ A_{HH} \cup A_{TH}, A_{HH} \cup A_{TT}, A_{HT} \cup A_{TH}, A_{HT} \cup A_{TT} \end{array}$$

11.2.3 Properties of probability measure

Based on the definition of probability space, we have following basic properties.

Lemma 11.2.1 (basic properties of probability measure). *[5, p. 11]*

- $P(\emptyset) = 0$.
- *(finite additivity) If* $A_1, A_2, ..., A_n \in \mathcal{F}$ *are* ***disjoints*** *subsets of* Ω*, we have:* $P(A_1 \cup A_2 ... \cup A_n) = P(A_1) + P(A_2) + ... + P(A_n)$.
- *For each* $A \in \mathcal{F}$, $P(A^C) = 1 - P(A)$*, where* A^C *is the complement of* A *with respect to* Ω.
- *If* $A_1, A_2 \in \mathcal{F}$ *and* $A_1 \subset A_2$*, then* $P(A_1) \leq P(A_2)$.
- *For* $B \subset A$, $P(A - B) = P(A) - P(B)$.
- *For* $A, B \in \mathcal{F}$*, we have*

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Proof. (1) Directly from

$$P(\cup\emptyset) = P(\emptyset) = \sum P(\emptyset)$$

and $P(\emptyset) \geq 0$, we have $P(\emptyset) = 0$. (2) Set $A_{n+1}, A_{n+2}, ... = \emptyset$ and use (1). (3) from (2). (4) note that $A_2 = A_1 + (A_2 - A_1)$ and $P(A_2 - A_1) \geq 0$. (5) Note that $(A - B) \cup B = A$ such that

$$P(B) + P(A - B) = P(A).$$

(6) Note that $A \cup B = A - A \cap B + B$ and the set $(A - A \cap B)$ and B are disjoint so that we can use (2) to prove. □

Example 11.2.2 (probability of drawing two cards from Poker cards).

- We first consider probability of drawing an ace or a King from Poker cards. Let A be the event that an ace is drawn and B the event that a King is drawn. It follows that $P(A) = \frac{4}{52} = \frac{1}{13}$ and $P(B) = \frac{4}{52} = \frac{1}{13}$. A and B are disjoint events. Then

$$P(A \cup B) = P(A) + P(B) = \frac{2}{13}.$$

- Now consider probability of drawing an ace or a spade from Poker cards. Let A be the event that an ace is drawn and B the event that a spade is drawn.

It follows that $P(A) = \frac{4}{52} = \frac{1}{13}$ and $P(B) = \frac{13}{52} = \frac{1}{4}$. A and B are not disjoint events. Then

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{1}{13} + \frac{1}{4} - \frac{1}{52} = \frac{4}{13}.$$

We can further derive a useful inequality, known as **union bound**.

Lemma 11.2.2 (union bound). *For any sequence $A_1, A_2, ... \in \mathcal{F}$ $P(A_1 \cup A_2 \cup A_3 \cup ...) \leq P(A_1) + P(A_2) + ...$*

Note that the equality holds when $A_1, A_2, ...$ are disjoint.

Proof. Based on countable additivity of probability function, we have:

$$\begin{aligned} P(A_1 \cup A_2 \cup A_3 \cup ...) &= P(A_1 \cup (A_2 \backslash A_1) \cup (A_3 \backslash A_2 \backslash A_1)...) \\ &= P(A_1) + P(A_2 \backslash A_1)... <= P(A_1) + P(A_2) + ... \end{aligned}$$

□

11.2.4 Conditional probability

11.2.4.1 *Basics*

In some random experiments, we are interested only in those outcomes that are elements of a subset C_1 of the sample space Ω. Then given the probability space $(\Omega, \mathcal{F}, P)$, and $C_1, C_2 \in \mathcal{F}$, the conditional probability of the event C_2, given C_1 is defined as

$$P(C_2|C_1) = \frac{P(C_1 \cap C_2)}{P(C_1)}.$$

Note that usually, for two events C_1, C_2 both occur, we can define an new event $C_3 = C_1 \cap C_2$, then we write $P(C_1, C_2) = P(C_3) = P(C_1 \cap C_2)$. $P(C_1 \cap C_2)$ is quite formal since it is based on set theory.

Definition 11.2.4 (conditional probability measure). *Given a probability space $(\Omega, \mathcal{F}, P)$ and an event $A \in \mathcal{F}, P(A) \neq 0$, we can define a conditional probability measure*

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

Lemma 11.2.3 (basic properties of conditional probability measure). *[5] Consider the conditional probability measure conditioned on event A. We have*

- $P(B|A) \geq 0, \forall B \in \mathcal{F}$.
- $P(B|A) = 0, \forall B \in \mathcal{F}, A \cap B$.
- $P(A|A) = 1$.
- $P(\cup_{j=1}^{\infty} B_i|A) = \sum_{j=1}^{\infty} P(B_i|A)$, *provided that* $B_1, B_2, ... \in \mathcal{F}$ *are mutually exclusive event.*
- $\sum_{i=1}^{\infty} P(C_i|A) = 1$, *where* $C_1, C_2, ... \in \mathcal{F}$ *are the partition of* Ω.

Proof. (4)Use countable additivity property of the definition of probability space Definition 11.2.2, we have

$$P(\cup_{j=1}^{\infty} B_i|A) = \frac{P(\cup_{j=1}^{\infty} B_i \cap A)}{P(A)} = \sum_{j=1}^{\infty} \frac{P(B_i \cap A)}{P(A)} = \sum_{j=1}^{\infty} P(B_i|A).$$

(5) Note that $\cup_{i=1}^{\infty}(C_i \cap A) = A$. □

Theorem 11.2.1 (Law of total probability). *Given a set of subsets $C_1, C_2, ..., C_k$, which are mutual disjoint and partition the sample space Ω, then we have*

$$P(C) = P(C \cap C_1) + P(C \cap C_2)... + P(C \cap C_k) = \sum_{i=1}^{k} P(C_i)P(C|C_i)$$

Proof. Note that we have $P(C \cap C_i) = P(C_i)P(C|C_i)$, then we get the law of total probability as:

$$P(C) = P(C_1)P(C|C_1) + P(C_2)P(C|C_2) + .. + P(C_k)P(C|C_k) = \sum_{i=1}^{k} P(C_i)P(C|C_i).$$

□

11.2.4.2 *Bayes' theorem*

The conditional probability formula also offers a convenient way to calculate intersection probabilities. Given events A and B, we have

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

Rearranging the formula and we get

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

By replacing $P(B)$ with the total probability law formula [Theorem 11.2.1], we have the following Bayes' theorem.

Theorem 11.2.2 (Bayes' theorem). *From the definition of the conditional probability, we have Bayes' theorem as:*

$$P(C_j|C) = \frac{P(C \cap C_j)}{P(C)} = \frac{P(C_i)P(C|C_i)}{\sum_{i=1}^{k} P(C_i)P(C|C_i)}$$

Proof. The law of total probability has been in the denominator. □

Example 11.2.3. Suppose that a diagonal test for some disease has following performance: for patients with this disease, the test produces a positive result with a probability of 98%; for patient without the disease, there would be a positive test result (i.e., false positive) with a probability 2%. If we randomly select a person, the probability of having the disease is 0.1%.

Now given a positive test result, what is the probability that the individual actually has the disease? Let A be the event that the individual has the disease and B be the event that the individual tests positive for the disease. Using Bayes' theorem the probability that a person who tests positive actually has the disease is

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

We have $P(A) = 1/1000, P(\overline{(A)}) = 1 - 1/1000,$; $P(B|A) = 98/100$ and $P(B|\overline{A}) = 2/100$. The rest of the calculation is straight forward.

11.2.4.3 *Independence of events and sigma algebra*

Definition 11.2.5 (independence of event). *Given the probability space $(\Omega, \mathcal{F}, P)$, and $C_1, C_2 \in \mathcal{F}$, then we say C_1 and C_2 are **independent** if*

$$P(C_1 \cap C_2) = P(C_1)P(C_2).$$

Definition 11.2.6 (independence of σ algebras). *Given the probability space $(\Omega, \mathcal{F}, P)$, and $\mathcal{F}_1, \mathcal{F}_2 \subset \mathcal{F}$, then we say $\mathcal{F}_1$ and $\mathcal{F}_2$ are **independent** if*

$$P(A_1 \cap A_2) = P(A_1)P(A_2), \forall A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2.$$

Remark 11.2.2. Note that this is mathematical equivalent definition, which does not reveal the nature of independence in terms of set relationship. **The nature is that if two events are independent, then the occurrence of one event will not change our brief on the occurrence of the other event.**

Example 11.2.4. Consider the sample space of a random experiment is given as $\{(0,0),(0,1),(1,0),(1,1)\}$, with its σ field consists of all its subsets, and we define event $C_1 = \{(0,0),(0,1)\}$, and $C_2\{(0,1)\}$. So the occurrence of C_1 will change the our brief of C_2 from $1/4$ to $1/2$. Therefore, C_1 and C_2 are not independent to each other. Also, consider $C_3 = C_1^c$, then the occurrence of C_1 change our brief of C_1^c to 0. If $C_4 = \Omega$, then the occurrence of C_4 will not change, and thus C_4 is always independent of other events. In summary, **independence between events is far more complicated than the simple set relations between events**

Remark 11.2.3. Here is an non-trivial example of independence. Consider the sample space as the product of two coin toss sample space, the event that the first toss get 1 is $\{(1,0),(1,1)\}$, which is independent of the other event that the second toss get 1(i.e. $\{(0,1),(1,1)\}$). The two events have finite intersections, but they are independent. Therefore, it seems that simply considering the set relationships between events can not yield complete information of independence. The nature of the random experiment, i.e., the probability measure, dictates the Independence. The intuition way to judge independence will be whether the occurrence of one events provides useful information, i.e., changes our brief, for the occurrence of the other event.

Lemma 11.2.4 (basic properties of independence). *[1]*

- *If $P(A) > 0$, then A and B are independent if and only if $P(B|A) = P(B)$*
- *If A and B are independent, then A and B^C are independent.*
- *If $P(A) = 0$ or 1, then for any $B \in \mathcal{F}, B \neq A$, A and B are independent.*
- *(independence of complements) If C_1, C_2 are independent, then C_1 and C_2^c,C_1^c and C_2, C_1^c and C_2^c are independent.*

Proof. (1) Suppose A and B are independent, then

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

(2)

$$P(A \cap B^C) = P(A \cap (\Omega - B)) = P(A \cap \Omega) - P(A \cap B) = P(A) - P(A)P(B) = P(A)P(B^C).$$

(3) If $A = \Omega$ such that $P(A) = 1$,then

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

If $A = \emptyset$ such that $P(A) = 0$,then

$$P(A \cap B) = P(A) = 0 = P(A)P(B).$$

□

11.3 Measurable map and random variable

11.3.1 Random variable

Definition 11.3.1 (measurable map). *[6] Let $(\Omega, \mathcal{F})$ and $(S, \mathcal{S})$ be two measurable space. A map $T : \Omega \to S$ is called $(\mathcal{F}, \mathcal{S})$-**measurable map** if*

$$T^{-1}(A) \in \mathcal{F}, \forall A \in \mathcal{S}$$

We can also write it as

$$T : (\Omega, \mathcal{F}) \to (S, \mathcal{S}).$$

Lemma 11.3.1 (basic properties of measurable map). *Let $(\Omega, \mathcal{F})$ and $(S, \mathcal{S})$ be two measurable space. Let a map $T : \Omega \to S$ be a measurable map. Then we have:*

- *For any two disjoint sets $S_1, S_2 \in \mathcal{S}$, $T^{-1}(S_1)$ and $T^{-1}(S_2)$ are disjoint.*
- *$T^{-1}(S) = \Omega$.*
- *(measurable composition preserves measurability) Let G be a measurable map from $(S, \mathcal{S})$ to $(S, \mathcal{S})$. Then $G \circ T : \Omega \to S$ is a measurable map.*

Proof. (1) Suppose their inverse image intersection M is nonempty, then $T(m), m \in M$ will map a single element to two different elements in S, which violates the definitions of mapping. (2) Suppose $T^{-1}(S) = \Omega_1 \subset \Omega$ and $\Omega_1 \neq \Omega$, then $T(\Omega - \Omega_1) = \emptyset$(otherwise T will map a single element to two different elements). Therefore $T^{-1}(S \cup \emptyset) = T^{-1}(S) = \Omega$. (3) Note that $(G \circ T)^{-1}(B) = T^{-1} \circ G^{-1}(B) = T^{-1}(G^{-1}(B))$ Because G is measurable map, $G^{-1}(B) \in \mathcal{S}$. Because T is measurable map, $T^{-1}(G^{-1}(B)) \in \mathcal{F}$. Therefore, $G \circ T$ is a measurable from Ω to S. □

Definition 11.3.2 (random variable in real space). *Let $(\Omega, \mathcal{F})$ and $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ be two measurable spaces defined on sample space Ω and $\mathbb{R}$, respectively.*

- *. A **random variable** in real space is a measurable map $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$.*
- *A **n-dimensional real-valued random vector** is a measurable map $X : (\Omega, \mathcal{F}) \to (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$.*

This following theorem provides the foundation of when X and Y are random variables, usually, $f(X), X + Y, XY, X/Y, ...$ are also random variables.

Theorem 11.3.1 (basic measurability properties of random variables). *Let $(\Omega, \mathcal{F})$ and $(\mathbb{R}, \mathcal{B})$ be two measurable space. Let $X : \Omega \to \mathbb{R}$ be a random variable. Then we have:*

- *(measurable composition preserves measurability) Let f be a measurable map from $(S, \mathcal{S})$ to $(S, \mathcal{S})$. Then $f(X) : \Omega \to \mathbb{R}$ is a measurable map.*
- *Let Y be another random variable from $(\Omega, \mathcal{F})$ to $(R, \mathcal{B})$. Then $\alpha X + \beta Y : \Omega \to \mathbb{R}, \alpha, \beta \in \mathbb{R}$ is also a measurable map.*
- *Let Y be another random variable from $(\Omega, \mathcal{F})$ to $(R, \mathcal{B})$. Then $XY : \Omega \to \mathbb{R}$ is also a measurable map.*
- *Let $Y, Y \neq 0$ be another random variable from $(\Omega, \mathcal{F})$ to $(R, \mathcal{B})$. Then $1/Y : \Omega \to \mathbb{R}$ is also a measurable map.*

Proof. (1) Use the composition property of measurable map [Lemma 11.3.1]. (2)(3)(4) use Lemma 3.8.4. □

When we define a random variable, we have also defined a new sample space (the range of the random variable) in $\mathbb{R}$. The original probability measure and the σ algebra on this new sample space (i.e., the Borel algebra) form a new probability space, as we show in the following theorem.

Theorem 11.3.2 (generation of probability space via random variable). *Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let $X : \Omega \to \mathbb{R}$ be a random variable. For each Borel set $B \in \mathcal{B}, B \subset \mathbb{R}$, we have $X^{-1}(B) \in \mathcal{F}$, then we can define $P_X(B) = P(X^{-1}(B))$. Then $(\mathbb{R}, \mathcal{B}, P_X)$ is a probability space.*

Proof. First $(\mathbb{R}, \mathcal{B})$ form a measurable space. So we only need to check the axiom property of P_X: (1) $P_X(A) \geq 0, \forall A \in \mathcal{B}$; (2) For any two disjoint sets A_1, A_2, then

$$P_X(A_1 \cup A_2) = P(X^{-1}(A_1 \cup A_2)) = P(X^{-1}(A_1) \cup X^{-1}(A_2)) = P(X^{-1}(A_1) + P(X^{-1}(A_2)).$$

We can directly generalize to countable additivity. (3) $P(X^{-1}(\mathbb{R})) = P(\Omega) = 1$ (from lemma on basic properties of measurable maps) □

This theorem paves the way for us **to directly work on this generated probability space $(\mathbb{R}, \mathcal{B}, P_X)$ and investigate distribution, density functions, etc., without referring back to the original probability space.**

11.3.2 σ algebra of random variables

Definition 11.3.3 (σ algebra generated by random variables). *[4, p. 52] Let X be a random variable map from nonempty Ω to $\mathbb{R}$. The σ**algebra generated** by X, denoted by $\sigma(X)$, is the collection of all subsets of Ω of the form $\{\omega \in \Omega : X(\omega) \in B\}$, or equivalently $X^{-1}(B)$, where B ranges over all Borel subsets of $\mathbb{R}$.*

Remark 11.3.1 (interpretation).

- When we define the measurable map from $(\Omega, \mathcal{F})$ to $(\mathbb{R}, \mathcal{B})$, usually $\sigma(X) \subseteq \mathcal{F}$. For example, if $X = const$, then $\sigma(X) = \mathcal{F}_0 = \{\emptyset, \Omega\}$.
- We cannot have $\mathcal{F} \subset \sigma(X), \mathcal{F} \neq \sigma(X)$ because the definition of random variable require measurability.

Definition 11.3.4 (measurable random variables with respect to a σ algebra). *Let X be a random variable map from nonempty Ω to $\mathbb{R}$. Let $\mathcal{G}$ be the σ algebra defined on Ω. We say X is $\mathcal{G}$ measurable if $\sigma(X) \subseteq \mathcal{G}$.*

Remark 11.3.2 (interpretation).

- Note that for any $B \in \mathcal{B}$, $X^{-1}(B) \in \sigma(X) \subseteq \mathcal{G}$, therefore X is also $\mathcal{G}$-measurable.
- Given a set Ω, we can define different σ algebra, including $\mathcal{F}_0 = \{\emptyset, \Omega\}$. But only σ algebra finer than $\sigma(X)$ can measure the mapping X.

Example 11.3.1.
- Consider a random experiment of tossing three dices. The sum of tossing outcomes is a random variable.
- In a random experiment of tossing coins 100 times. The total number of heads is a random variable.

11.3.3 Independence of random variables

Definition 11.3.5 (independence of random variables). *Let $X : (\Omega, \mathcal{F}) \to (S, \mathcal{S})$ and $Y : (\Omega, \mathcal{F}) \to (S, \mathcal{S}))$ denote two random variables. We say X, Y are independent, if for all $A, B \in \mathcal{S}$ the events $X^{-1}(A)$ and $Y^{-1}(B)$ are independent in the sense that $P(X^{-1}(A) \cap Y^{-1}(B)) = P(X^{-1}(A))P(Y^{-1}(B))$.*

Definition 11.3.6 (independence of random variables, alternative). *Let* $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ *and* $Y : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ *denote two random variables. We say* X, Y *are independent, if for any events* $A, B, A \in \sigma(X), B \in \sigma(Y)$

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

Remark 11.3.3. Note that

- independence of random variables are much more than independence of a selected set of events, because it requires that *all* events are independent to each other.
- if X, Y are map from different sample space, then they are independent.

Lemma 11.3.2 (function composition preserves random variable independence). *Let* X, Y *be independent random variables defined from* Ω *to* $\mathbb{R}$ *, and let* f *and* g *be Borel-measurable functions on* $\mathbb{R}$*. Then* $f(X)$ *and* $g(Y)$ *are independent random variables.*

Proof. Note that for any $B \in \mathcal{B}$, $f^{-1}(B) \in \mathcal{B}$ since f is Borel measurable. Then $X^{-1}(f^{-1}(B)) \in \sigma(X)$ based on the definition of σ generation. Therefore, $\sigma(f(X)) \subset \sigma(X)$. Similarly, $\sigma(g(Y)) \subset \sigma(Y)$. Since every events in $\sigma(X)$ and $\sigma(Y)$ are independent, then every events in $\sigma(f(X))$ and $\sigma(g(Y))$ are independent; that is, $f(X)$ and $g(Y)$ are independent random variables. □

11.4 Distributions of random variables

11.4.1 Basic concepts

11.4.1.1 *Probability mass function*

Definition 11.4.1 (random variable, random vector).

- *Let X be a random variables maps from the probability space $(\Omega, \mathcal{F}, P)$ to $\mathbb{R}$. The **space** of the random variable X is the set*

$$\{(X(\omega) : \omega \in \Omega\}.$$

- *Let $X_1, X_2, ..., X_n$ be random variables maps from the probability space $(\Omega, \mathcal{F}, P)$ to $\mathbb{R}$. We say $(X_1, X_2, ..., X_n)$ is a random vector. The **space** of the random vector $(X_1, X_2, ..., X_n)$ is the set*

$$\{(X_1(\omega), X_2(\omega), ..., X_n(\omega)) : \omega \in \Omega\}.$$

Definition 11.4.2 (probability mass function, pmf).

- *For a discrete random variable X with space $\mathcal{D}$, the **probability mass function (pmf)** to characterize its distribution is given by*

$$f_X(x) = P(X = x), \forall x \in \mathcal{D}.$$

- *For a discrete random vector $(X_1, X_2, ..., X_n)$ with space $\mathcal{D}$, the **joint probability mass function** to characterize its distribution is given by*

$$f_{X_1, X_2, ..., X_n}(x_1, x_2, ..., x_n) = P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n), \forall (x_1, x_2, ..., x_n) \in \mathcal{D}.$$

Example 11.4.1. Consider the experiment of toss a biased coin. The probability of getting head is p and getting tail is $1 - p$. The outcome of coin toss can be modeled by a Bernoulli random variable X, whose pmf is given by

$$f_X(x) = p^x(1 - p)^{1-x}, x \in \{0, 1\}.$$

11.4.1.2 *Distributions on $\mathbb{R}^n$*

Definition 11.4.3 (cumulative distribution functions).

- *Let X be a random variable with space $\mathcal{D} \subset \mathbb{R}$. The cumulative distribution function for X is given by*
$$F_X(x) = P(X \leq x).$$
- *Let $(X_1, X_2, ..., X_n)$ be a random vector with space $\mathcal{D} \subset \mathbb{R}^n$. The joint cumulative distribution function for X is given by*
$$F_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n) = P(X_1 \leq x_1, X_2 \leq x_2, ..., X_n \leq x_n).$$

Remark 11.4.1 (A rigorous interpretation). [6]

- Let P denotes a probability measure of the original probability space $(\Omega, \mathcal{F}, P)$. Let X be the random variable, then
$$F_X(x) := P(X \leq x) = P(X^{-1}((-\infty, x))).$$
where X^{-1} maps a measurable subset in $\mathcal{B}(\mathbb{R})$ to a measurable set in $\mathcal{F}$.
- Note that every subset of such form $(-\infty, x), x \in \mathbb{R}$ is a member of $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and therefore $P(X \leq x) = P(X^{-1}((-\infty, x)))$ has a well-defined value.

Definition 11.4.4 (marginal cdf). *Let $(X_1, X_2, ..., X_n)$ be a random vector with joint cdf $F_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n)$. The marginal cdf of X_i is defined by*
$$F_{X_i}(x_i) = P(X_1 < \infty, ..., X_i \leq x_i, ..., X_n < \infty).$$

Lemma 11.4.1 (area probability formula). *Let X_1, X_2 be random variables with joint cdf $F_{X_1,X_2}(x_1, x_2)$. Then*
$$\begin{aligned} &P(a_1 \leq X_1 \leq b_1, a_2 \leq X_2 \leq b_2) \\ =&F_{X_1,X_2}(b_1, b_2) - F_{X_1,X_2}(a_1, b_2) - F_{X_1,X_2}(b_1, a_2) + F_{X_1,X_2}(a_1, a_2) \end{aligned}$$

Proof. Straight forward. □

11.4.1.3 *Probability density function*

Definition 11.4.5 (probability density function, pdf).

- *Let X be a random variable with cdf $F_X(x)$. The probability density function for X is defined by*
$$f_X(x) = \frac{dF_X(x)}{dx}.$$
- *Let $(X_1, X_2, ..., X_n)$ be a random vector with joint cdf $F_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n)$. The joint probability density function for $(X_1, X_2, ..., X_n)$ is given by*
$$F_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n) = \frac{\partial^n F_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n)}{\partial x_1 \partial x_2 \cdots \partial x_n}.$$

Definition 11.4.6 (support of a random variable). *Let X be a random variable with pdf f_X and space $\mathcal{D}$. The support of X is defined as the set*
$$S_X = \{x \in \mathcal{D} : f_X(x) > 0\}.$$

Example 11.4.2. A random variable X with normal distribution $N(\mu, \sigma^2)$, characterized by parameters μ and σ, has its pdf given by
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2}(x-\mu)^2/\sigma^2), -\infty < x < \infty.$$
The support of X is $\mathbb{R}$.

Definition 11.4.7 (marginal pdf). *Let $(X_1, X_2, ..., X_n)$ be a random vector with joint pdf $f_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n)$. The marginal pdf of X_i is defined by*
$$f_{X_i}(x) = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} f_{X_1,X_2,...,X_n} dx_1 ... dx_{i-1} dx_{i+1} ... dx_n.$$
If the marginal cdf of X_i is F_{X_i}, then
$$f_{X_i}(x) = \frac{dF_{X_i}(x)}{dx}.$$

11.4.1.4 *Conditional distributions*

Definition 11.4.8 (conditional probability mass function (pmf)). *Let X_1 and X_2 be discrete random variables with joint pmf $p_{X_1,X_2}(x_1,x_2)$. Let $p_{X_1}(x_1)$ denote the marginal pmf Let x_1 be a point such that $p_{X_1}(x_1) > 0$.*

The conditional pmf of X_2 given $X_1 = x_1$ is defined as

$$p_{X_2|X_1}(x_2|x_1) = \frac{p_{X_1,X_2}(x_1,x_2)}{p_{X_1}(x_1)}.$$

Remark 11.4.2. The sum to 1 property can be verified by

$$\begin{aligned}
&\sum_{x_2} p_{X_2|X_1}(x_2|x_1) \\
&= \sum_{x_2} \frac{p_{X_1,X_2}(x_1,x_2)}{p_{X_1}(x_1)} \\
&= \frac{1}{p_{X_1}(x_1)} \sum_{x_2} p_{X_1,X_2}(x_1,x_2) = \frac{p_{X_1}(x_1)}{p_{X_1}(x_1)} = 1
\end{aligned}$$

Definition 11.4.9 (conditional probability density function (pdf)). *Let X_1 and X_2 be discrete random variables with joint pdf $f_{X_1,X_2}(x_1,x_2)$. Let $f_{X_1}(x_1)$ denote the marginal pmf Let x_1 be a point such that $f_{X_1}(x_1) > 0$.*

The conditional pdf of X_2 given $X_1 = x_1$ is defined as

$$f_{X_2|X_1}(x_2;x_1) = \frac{f_{X_1,X_2}(x_1,x_2)}{f_{X_1}(x_1)}.$$

Lemma 11.4.2 (basic properties). *[5, p. 97]*

- $f_{X_2|X_1}(x_2;x_1) > 0$
- $\int_{-\infty}^{\infty} f_{X_2|X_1}(x_2;x_1)dx_2 = 1.$
- $\int_{-\infty}^{\infty} f_{X_2|X_1}(x_2;x_1)f_{X_1}(x_1)dx_1 = f_{X_2}(x_2).$
- $E[u(X_2)|x_1] = \int_{-\infty}^{\infty} f_{X_2|X_1}(x_2;x_1)u(x_1)dx_2$

11.4.1.5 *Bayes law*

Theorem 11.4.1 (Bayes law for random variables). *Let X, Y, Z be random variables. It follows that*

- *(unconditional Bayesian law)*

$$f(X|Y) = \frac{f(Y|X)f(X)}{\int f(Y|X)f(X)dx}$$

- *(conditional Bayesian law)*

$$f(X|Y,Z) = \frac{f(X|Z)f(Y|X)}{\int f(Y|X)f(X|Z)dx}$$

Proof. (1) Note that the denominator $\int f(Y|X)f(X)dx = f(Y)$. Therefore

$$f(X|Y)\int f(Y|X)f(X)dx = f(X,Y) = f(Y|X)f(X).$$

(2) Note that

$$\int f(Y|X)f(X|Z)dx = \int f(Y,X|Z)dx = f(Y|Z).$$

Therefore,

$$f(X|Y,Z)\int f(Y|X)f(X|Z)dx = f(X|Y,Z)f(Y|Z) = f(X,Y|Z).$$

□

11.4.2 Independence

Definition 11.4.10 (independence of random variables). *[5, pp. 112, 115] Let the random variables X_1 and X_2 have joint pdf $f(x_1,x_2)$ and marginal pdfs $f_1(x_1), f_2(x_2)$.*

- *The random variables X_1 and X_2 are said to be independent if and only*

$$P(a < X_1 \leq b, c < X_2 \leq d) = P(a < X_1 \leq b, c < X_2 \leq d),$$

  for every $a < b, c < d$, where a, b, c, d are constants.
- *The random variables X_1 and X_2 are said to be independent if and only*

$$f(x_1,x_2) = f_1(x_1)f_2(x_2).$$

Lemma 11.4.3 (conditions for independence). *[5, p. 113] Let the random variables X_1 and X_2 have supports S_1 and S_2, and have joint pdf $f(x_1, x_2)$. Then X_1 and X_2 are independent if and only if $f(x_1, x_2)$ can be written as*

$$f(x_1, x_2) = g(x_1)h(x_2),$$

where $g(x_1) > 0, x_1 \in S_1$, zero elsewhere, and $h(x_2) > 0, x_2 \in S_2$, zero elsewhere.

Proof. (1) If $f(x_1, x_2) = g(x_1)h(x_2)$, then

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} g(x_1)h(x_2)dx_2 = c_2 g(x_1).$$

$$f_{X_2}(x_1) = \int_{-\infty}^{\infty} g(x_1)h(x_2)dx_1 = c_1 h(x_2).$$

Further we have

$$1 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x_1)h(x_2)dx_1dx_2 = c_1c_2.$$

Therefore, $f(x_1, x_2) = c_1c_2g(x_1)h(x_2) = f_{X_1}(x_1)f_{X_2}(x_2)$; that is, X_1, X_2 are independent.

(2) The other direction directly from definition. □

Lemma 11.4.4 (independence criterion from cdf). *[5, p. 114] Let the random variables X_1 and X_2 have joint cdf $F(x_1, x_2)$ and marginal cdfs $F_1(x_1), F_2(x_2)$. Then X_1 and X_2 are independent if and only if*

$$F(x_1, x_2) = F_1(x_1)F_2(x_2).$$

Proof. (1) From Lemma 11.4.1, we have

$$\begin{aligned}
&P(a_1 \le X_1 \le b_1, a_2 \le X_2 \le b_2)\\
=&F_{X_1,X_2}(b_1, b_2) - F_{X_1,X_2}(a_1, b_2) - F_{X_1,X_2}(b_1, a_2) + F_{X_1,X_2}(a_1, a_2)\\
=&F_{X_1}(b_1)F_{X_2}(b_2) - F_{X_1}(a_1)F_{X_2}(b_2) - F_{X_1}(a_2)F_{X_2}(b_1) + F_{X_1}(a_1)F_{X_2}(a_2)\\
=&(F_{X_1}(b_1) - F_{X_1}(a_1))(F_{X_2}(b_2) - F_{X_2}(a_2))\\
=&P(a_1 \le X_1 \le b_1)P(a_2 \le X_2 \le b_2)
\end{aligned}$$

Since a_1, a_2, b_1, b_2 are arbitrary, X_1 and X_2 are independent. (2)The other direction directly from definition. □

11.4.3 Conditional independence

Definition 11.4.11 (conditional independence). *Given discrete random variables X, Y, and Z, we say X and Y are conditionally independent on Z if we can be write:*

$$P(X, Y|Z = z) = P(X|Z = z)P(Y|Z = z).$$

If not conditionally independent, we will have

$$P(X, Y|Z = z) = P(X|Y, Z = z)P(Y|Z = z).$$

Remark 11.4.3. Intuitively, two random variable X, Y are conditional independence given Z is that: if the value of Z is known, X, Y are independent to each other,i.e., the occurrence of events about Y will not give extra information to the occurrence of events abouts X. We need to distinguish two different cases:

- If X, Y are independent, then they are conditionally independent to each other.
- If events about Z already gives information contained in events about Y, then X, Y are conditionally independent given Z.

Remark 11.4.4. Conditionally independence will help us simplify calculation, for example:

$$P(X = x|Y = y, Z = z) = P(X = x|Z = z)$$

if X, Y are conditionally independent given Z.

11.4.4 Transformations

11.4.4.1 *Transformation for univariate distribution*

Lemma 11.4.5 (change of variable). *Let X have cdf $F_X(x)$ and Let $Y = g(X)$, where g is a **monotonely increasing** function. Then,*

$$F_Y(y) = F_X(g^{-1}(y)).$$

*If g is a **monotonely decreasing function**, then*

$$F_Y(y) = 1 - F_X(g^{-1}(y)).$$

Proof. If g is increasing function

$$P(Y < y) = P(g(X) < y) = P(X < g^{-1}(y)) = F_X(g^{-1}(y)).$$

If g is decreasing function

$$P(Y < y) = P(g(X) < y) = P(X > g^{-1}(y)) = 1 - F_X(g^{-1}(y)).$$

□

Lemma 11.4.6 (change of variable). *Let X have pdf $f_X(x)$ and Let $Y = g(X)$, where g is a **monotone** function. Let X and Y be defined as*

$$\mathcal{X} = \{x : f_X(x) > 0\}, \mathcal{Y} = \{y : y = g(x), x \in \mathcal{X}\}$$

then we have

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|, y \in \mathcal{Y} \\ 0, otherwise \end{cases}$$

Remark 11.4.5 (why monotonicity). We require the $g(X)$ to be monotone because if $g'(x)$ has different sign on different regions, then $g'(x_0) = 0$ for some x_0 and $g(x)$ is not invertible near the neighborhood of x_0.

11.4.4.2 *Location-scale transformation*

Lemma 11.4.7 (location-scale transformation). *Demote the pdf and cdf of a random variable Z as f_Z and F_Z. Then for any $\mu, \sigma \in \mathbb{R}, \sigma > 0$, we have:*

- *The random variable $X = \sigma Z + \mu$ is a random variable with pdf*

$$F_X(x) = F_Z(\frac{x - \mu}{\sigma}).$$

- $$f_X(x) = \frac{1}{\sigma} f_Z(\frac{x - \mu}{\sigma}).$$

- *The change of percentile. $F_X^{-1}(\alpha) = \mu + \sigma F_Z^{-1}(\alpha), \forall \alpha \in [0, 1]$, where $F_X^{-1}(\alpha) = \inf\{P(X < x) \geq \alpha\}$.*

Proof. For (1)(2)

$$\begin{aligned} F_X(x) &= P(X < x) \\ &= P(\mu + \sigma Z < x) \\ &= P(Z < (x - \mu)/\sigma) \\ &= F_Z((x - \mu)/\sigma) \end{aligned}$$

Then

$$f_X(x) = dF_X(x)/dx = \frac{1}{\sigma} f_Z(\frac{x-\mu}{\sigma}).$$

(3)

$$\begin{aligned} \alpha &= P(X < F_X^{-1}(\alpha)) \\ &= P(\sigma Z + \mu < F_X^{-1}(\alpha)) \\ &= P(Z < (F_X^{-1}(\alpha) - \mu)/\sigma) \\ \alpha &= F_Z((F_X^{-1}(\alpha) - \mu)/\sigma) \\ \implies F_Z^{-1}(\alpha) &= (F_X^{-1}(\alpha) - \mu)/\sigma \\ \mu + \sigma F_Z^{-1}(\alpha) &= F_X^{-1}(\alpha). \end{aligned}$$

□

Example 11.4.3. Consider the random variable $X \sim N(0,1)$ with

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2}).$$

Let $Y = \sigma X + \mu$, then

$$f_Y(y) = \frac{1}{\sigma} f_X(\frac{y-\mu}{\sigma}) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(x-\mu)^2}{2\sigma^2}).$$

11.4.4.3 *Transformation for multivariate distribution*

Lemma 11.4.8 (multivariate transformation). *[5, p. 128] Let $(X_1, X_2, ..., X_n)$ be a random vector with support $\mathcal{S}$. Let*

$$y_1 = y_1(x_1, ..., x_n), ..., y_n = y_n(x_1, ..., x_n)$$

define a set of transformations with inverse

$$x_1 = x_1(y_1, ..., y_n), ..., x_n = x_n(y_1, ..., y_n).$$

Let $\mathcal{T}$ be the image of $\mathcal{S}$ under the transformation.

Let $f_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n)$ be the joint pdf of $(X_1, X_2, ..., X_n)$. Then the joint pdf for the random vector $(Y_1, Y_2, ..., Y_n)$ is given by

$$f_{X_1,X_2,...,X_n}(x_1,x_2,...,x_n) = f_{Y_1,Y_2,...,Y_n}(y_1(x_1,x_2,...,x_n),...,y_n(x_1,x_2,...,x_n))|J|$$

or

$$f_{X_1,X_2,...,X_n}(x_1(y_1,y_2,...,y_n),...,x_n(y_1,y_2,...,y_n)) = f_{Y_1,Y_2,...,Y_n}(y_1,y_2,...,y_n)|J|$$

where

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \dots & \frac{\partial x_1}{\partial y_n} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & & \frac{\partial x_2}{\partial y_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial x_n}{\partial y_1} & \frac{\partial x_n}{\partial y_2} & \dots & \frac{\partial x_n}{\partial y_n} \end{vmatrix}.$$

Moreover,

$$\int_{\mathcal{T}} f_{Y_1,...,Y_n}(y_1,...,y_n)|J|\,dy_1...dy_n = 1$$

Proof. (1)For S be a measurable subset in $\mathcal{S}$, let $T \in \mathcal{T}$ denote the its image under the transformation. We have

$$P((Y_1,Y_2,...,Y_n) \in T) = P((X_1,X_2,...,X_n) \in S)$$

Note that

$$P((X_1,X_2,...,X_n) \in S) = \int_S f_{X_1,...,X_n}(x_1,...,x_n)dx_1...dx_n$$

and

$$dx_1dx_2 = |J|\,dy_1dy_2.$$

Then

$$\int_S f_{X_1,...,X_n}(x_1,...,x_n)dx_1...dx_n = \int_T f_{Y_1,...,Y_n}(y_1(x_1,...,x_n),...,y_n(x_1,...,x_n))|J|\,dy_1...dy_n.$$

Because S is arbitrary, we have

$$f_{X_1,...,X_n}(x_1,...,x_n) = f_{Y_1,...,Y_n}(y_1(x_1,...,x_n),...,y_n(x_1,...,x_n))|J|\,.$$

(2)

$$\int_{\mathcal{T}} f_{Y_1,...,Y_n}(y_1,...,y_n)|J|\,dy_1...dy_n = \int_{\mathcal{S}} f_{X_1,...,X_n}(x_1,...,x_n)dx_1...dx_n = 1$$

□

Remark 11.4.6.

- We interpret dx_1dx_2 as infinitesimal area in the original $\mathcal{S}$, and this area is mapped to an area in $\mathcal{T}$. Note that we divide $\mathcal{S}$ and $\mathcal{T}$ into the same number small areas and the sum them up to calculate the integral. The areas in both $\mathcal{T}$ and $\mathcal{S}$ have the following relation:

$$dx_1dx_2 = |J|\, dy_1(x_1, ..., x_2)dy_2(x_1, ..., x_n) = |J|\, dy_1dy_2.$$

- If we maps from larger support to a smaller support, for example, from $\mathbb{R}^2$ to $[0,\infty) \times [0, 2\pi]$, the density will increase.
-

Lemma 11.4.9 (polar transformation). *Let (X_1, X_2) be a random vector with support $S = \mathbb{R}^2$. Let $R = \sqrt{X_1^2 + X_2^2}, \Theta = \arctan(X_1/X_2)$. Then*

- $$f_{R,\Theta}(r,\theta)r = f_{X_1,X_2}(r\cos(\theta), r\sin(\theta)).$$
- $$f_{X_1,X_2}(x_1,x_2)dx_1dx_2 = f_{R,\Theta}(r(x_1,x_2),\theta(x_1,x_2))rdrd\theta.$$
- $$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X_1,X_2}(x_1,x_2)dx_1dx_2 = \int_0^{\infty}\int_0^{2\pi} f_{R,\Theta}(r,\theta)rdrd\theta = 1.$$
- *The support for (R,Θ) is*
$$\{(0,+\infty)\times[0,2\pi]\}$$

Proof. Note that

$$J = \begin{vmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{vmatrix} = \begin{vmatrix} \cos(\theta) & -r\sin(\theta) \\ \sin(\theta) & r\cos(\theta) \end{vmatrix} \implies |J| = r.$$

Therefore

$$f_{X_1,X_2}(x_1,x_2) = f_{R,\Theta}(r(x_1,x_2),\theta(x_1,x_2))r.$$

□

Example 11.4.4. Let $X \sim N(0,1)$ and $Y \sim N(0,1)$. Let $R = \sqrt{X^2+Y^2}, \Theta = \arctan(X/Y)$.

Then

$$f_{X,Y}(x,y) = \frac{1}{2\pi}\exp(-\frac{x^2+y^2}{2}).$$

$$f_{R,\Theta}(r,\theta) = f_{X,Y}(r\cos(\theta), r\sin(\theta)) = \frac{1}{2\pi}\exp(-\frac{r^2}{2}).$$

$$\int_0^\infty \int_0^{2\pi} f_{R,\Theta}(r,\theta) r dr d\theta = \int_0^\infty \int_0^{2\pi} \frac{1}{2\pi}\exp(-\frac{r^2}{2}) r dr d\theta = 1.$$

Lemma 11.4.10 (convolution formula). *[5, p. 95] Let X_1 and X_2 be continuous random variables with joint pdf $f_{X_1,X_2}(x_1,x_2)$ with $\mathcal{D} = \mathbb{R}^2$. Let $Y_1 = X_1 + X_2$ and $Y_2 = X_2$. Then*

- $f_{Y_1,Y_2}(y_1,y_2) = f_{X_1,X_2}(y_1 - y_2, y_2)$.
- *The pdf of Y_1 is given by*

$$f_{Y_1}(y) = \int_{-\infty}^{\infty} f_{X_1,X_2}(y - y_2, y_2) dy dy_2.$$

Proof. It is easy to see $|J| = 1$. □

11.5 Expectation, variance, and covariance

11.5.1 Expectation

The expectation of a random variable is approximately the mean values averaging over a large number of random outcomes. Formally, we define the expectation using probability density function (pdf) and probability mass function (pmf) for continuous and discrete random variables, respectively.

Definition 11.5.1.

- *Let X be a continuous random variable with pdf $f_X(x)$. The expectation of X is*

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

- *Let X be a discrete random variable with pmf $f_X(x)$. The expectation of X is*

$$E[X] = \sum_x x f_X(x) = \sum_x x \mathbb{P}(X = x).$$

Let $g(X)$ be a function of a random variable X. Intuitively, the mean value of $g(X)$ are just the average over transformed random outcomes of X. Alternatively, we can view $g(X)$ as a new random variable under rather mild condition of g (i.e., measurability,).

Definition 11.5.2.

- *Let X be a continuous random variable, and let g be a function. The expectation of $g(X)$ is*

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

- *Let X be a discrete random variable, and let g be a function. The expectation of $g(\bar{X})$ is*

$$E[g(X)] = \sum_x g(x) f_X(x) = \sum_x g(x) P(X = x).$$

Remark 11.5.1 (probability as an expectation). Let A be any event, we can also express $P(A)$ as an expectation by defining a indicator random variable

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

and evaluate the expectation of I_A. We have Then I_A is a random variable, and

$$\begin{aligned} E[I_A] &= \sum_{r=0}^{1} rP\left(I_A = r\right) \\ &= 0 \times P\left(I_A = 0\right) + 1 \times P\left(I_A = 1\right) \\ &= P\left(I_A = 1\right) \\ &= P(A) \end{aligned}$$

The probability to expectation conversion allow generalization of theorems on probabilities to theorems on expectations.

11.5.2 Expectation in the Lebesgue framework

Let X be a random variable defined on a probability space $(\Omega, \mathcal{F}, P)$, if Ω is finite, we can simply define the expectation as

$$E[X] = \sum_{\omega \in \Omega} P(\omega) X(\omega)$$

However, if Ω is countably infinite, we can still list a sequence of $\omega_1, \omega_2, ...$ such that

$$E[X] = \sum_{i=1}^{\infty} P(\omega) X(\omega_i)$$

However, if Ω is uncountably infinite, then **uncountable** summation is not defined, and we need Lebesgue integral.

Definition 11.5.3 (Lebesgue integral). *[4, p. 15] Let X be a random variable defined on a probability space $(\Omega, \mathcal{F}, P)$, assume $0 \leq X(\omega) \leq \infty$ for every $\omega \in \Omega$, and let $\Pi : 0 = y_0 < y_1 < ...$ be a partition on the range of $X(\omega)$. For each subinterval $[y_k, y_{k+1}]$, we set*

$$A_k = \{\omega \in \Omega : y_k \leq X(\omega) \leq y_{k+1}\} = X^{-1}([y_k, y_{k+1}])$$

We define the lower Lebesgue sum to be

$$LS_{\Pi}^{-} = \sum_{k=1}^{\infty} y_k P(A_k)$$

We further define the limit

$$\lim_{\|\Pi\|\to 0} LS^{-}_{\Pi} = \int_{\Omega} X(\omega)dP(\omega)$$

Remark 11.5.2.

- Because X is measurable maps, its inverse image of any Borel set in $\mathbb{R}$ is measurable, i.e., $P(A)$ has value.
- For $X(\omega)$ that takes positive and negative part, we can simply decompose into two parts and use the linearity.

Definition 11.5.4 (expectation). *Let X be a random variable on a probability space $(\Omega, \mathbb{F}, P)$. The expectation of X is defined to be*

$$E[X] = \int_{\Omega} X(\omega)dP(\omega)$$

This definition makes sense if X is integrable, i.e., if

$$E|X| = \int_{\Omega} \left|X(\omega)\right| dP(\omega) < \infty$$

Remark 11.5.3. Note that the integral is defined using Lebesgue integral, and based on this definition we can recover the elementary definitions.

- If X takes only finitely many $x_0, x_1, ..., x_n$, but Ω is uncountable, then

$$EX = \sum_{x_k} x_k P(X = x_k)$$

  and $P(X = x_k)$ is the probability measure of all the subsets $X^{-1}(\{x_k\})$
- In particular, if Ω is finite, then

$$EX = \sum_{\omega\in\Omega} X(\omega)P(\omega)$$

Example 11.5.1. Let $\Omega = [0,1]$, and let P be the Lebesgue measure on $[0,1]$. Consider $X(\omega) = 1$, if ω is irrational; o otherwise. Then $E[X] = 1P(\omega \in [0,1] : \omega \text{ is irrational}) + 0P(\omega \in [0,1] : \omega \text{ is rational}) = 1$ since $P(\omega \in [0,1] : \omega \text{ is irrational}) = 1 = 1$, $P(\omega \in [0,1] : \omega \text{ is rational}) = 0$

Definition 11.5.5 (expectation of function of random variable). *Let $h : \mathbb{R} \to \mathbb{R}, f \in \mathcal{C}$, and let $X : \Omega \to \mathbb{R}$ be a random variable with probability density $f(x)$. Then the expectation of $h(X) : \Omega \to \mathbb{R}$ is given as:*

$$E[h(X)] = \int_{-\infty}^{+\infty} h(x)f(x)dx$$

11.5.3 Properties of expectation

Linearity of expectation is most fundamental property. It applies to random variables no matter they are independent or not. Linearity of expectation is the direct result of linearity of Lebesgue integral.

Lemma 11.5.1 (linearity of expectation). *Let X, Y be two random variables over the same probability space. Then*

$$E[\alpha X + \beta Y] = \alpha E[X] + \beta E[Y],$$

and

$$E[aX + b] = aE[X] + b,$$

where a and b are real-valued constants.

Lemma 11.5.2 (expectation and independence). *Let X, Y be two random variables and $g(\cdot)$ and $h(\cdot)$ be two functions. If X and Y are independent, then*

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

11.5.4 Variance and covariance

Definition 11.5.6 (variance, covariance). *The variance of a random variable X is defined as*

$$Var[X] = E[(X - EX)^2]$$

The covariance of random variables X and Y is defined by

$$Cov(X, Y) = E[(X - EX)(Y - EY)].$$

The covariance matrix $Cov(Z)$ of a random vector $Z = [Z_1, ..., Z_m]^T$ is defined as

$$Cov(Z)_{ij} = Cov(Z_i, Z_j).$$

Here we list a number of basic properties of variance and covariance. Most of them are straight forward or can be proved via linearity of expectation.

Lemma 11.5.3 (basic properties for variance and covariance). *Let X and Y be random variables, let $a, b \in \mathbb{R}$*

- $Var[X] = E[X^2] - E[X]^2$
- $Cov(X, Y) = E[XY] - E[X]E[Y]$
- $Cov(\sum_i^m a_i X_i, \sum_j^n b_j Y_j) = \sum_i^m \sum_j^n a_i b_j Cov(X_i, Y_j)$
- $Var[X + a] = Var[X]$
- $Var[aX] = a^2 Var[X]$
- $Var[aX + bY] = a^2 Var[X] + b^2 Var[Y] + 2abCov(X, Y)$
- $Var[aX - bY] = a^2 Var[X] + b^2 Var[Y] - 2abCov(X, Y)$
- *More generally,*

$$Var[\sum_{i=1}^n a_i X_i] = \sum_{i=1}^n a_i^2 Var[X_i] + 2\sum_{i=1}^n \sum_{j>1}^n a_i a_j Cov(X_i, X_j) = \sum_{i=1}^n \sum_{j=1}^n a_i a_j Cov(X_i, X_j)$$

Lemma 11.5.4 (basic properties for random vectors). *Let X be a random vector, let A, B be non-random matrices, we have*

- $Cov(AX) = ACov(X)A^T$
- $Cov(X + B) = Cov(X)$

Lemma 11.5.5 (variance of a function of a random variable). *Let X be a random variable taking value in $\mathcal{X}$ with pdf $f(x)$, let g be a continuous function, then*

$$Var[g(X)] = E[(g(X) - E[g(X)])^2]] = \int_{\mathcal{X}} (g(x) - E[g(x)])^2 f(x) dx$$

Proof. Note that $E[g(X)]$ is a constant. We can calculate $Var[g(X)]$ using the expectation of a function of a random variable definition [Definition 11.5.5]. □

11.5.5 Conditional variance

Theorem 11.5.1 (conditional variance identity). *[7, p. 193] For any two random variables X and Y,*

$$Var[X] = E[Var[X|Y]] + Var[E[X|Y]],$$

provided that the expectation exists.

Example 11.5.2. Suppose the random variable $Y \sim Binomial(n, X)$, where $X \sim Uniform(0, 1)$ and n is a given constant. Then we can calculate

$$E[Y] = E[E[Y|X]] = E[nX]$$

and

$$Var[Y] = Var[E[Y|X]] + E[Var[Y|X]] = Var[nX] + E[nX(1-X)].$$

11.5.6 Delta method

With the knowledge of mean and variance of a random variable, we can approximate the mean and variance of a function a random variable via Taylor expansion.

Lemma 11.5.6 (first-order approximation to mean and variance of a function). *[7, p. 242] Let $X_1, ..., X_k$ be random variables with mean $\mu_1, ..., \mu_k$, and define $X = (X_1, ..., X_k)$ and $\mu = (\mu_1, ..., \mu_k)$. Define a differentiable function $g : \mathbb{R}^k \to \mathbb{R}$. Xhen we have the following first-order approximate mean and variance:*

$$\begin{aligned} E[g(X)] &\approx g(\mu) \\ Var[g(X)] &\approx \sum_{i=1}^{k} [g_i'(\mu)]^2 Var[X_i] + 2 \sum_{i=1}^{k} \sum_{j>i}^{k} g_i'(\mu) g_j'(\mu) Cov_{ij}[X]. \end{aligned}$$

Proof. (1)

$$\begin{aligned} g(X = t) &= g(\mu) + \sum_{i=1}^{k} g_i'(\mu)(t - \mu) + o((t - \mu)) \\ &\approx g(\mu) + \sum_{i=1}^{k} g_i'(\mu)(t - \mu) \\ E[g(X)] &\approx g(\mu) + \sum_{i=1}^{k} g_i'(\mu)(E[X] - \mu) = g(\mu) \end{aligned}$$

(2)

$$g(X = t) \approx g(\mu) + \sum_{i=1}^{k} g_i'(\mu)(t - \mu)$$
$$Var[g(X)] \approx Var[\sum_{i=1}^{k} g_i'(\mu)(X - \mu)] = \sum_{i,j} g_i'(\mu) g_j'(\mu) Cov_{ij}$$

□

Corollary 11.5.1.1. *Let $X_1, ..., X_k$ be iid random samples of X. Assume $E[X] = \mu$. Let $g : \mathbb{R} \to \mathbb{R}$ be a differentiable function. Xhen, we have the first-order approximation:*

$$E[g(X)] \approx g(\mu)$$
$$Var[g(X)] \approx [g'(\mu)]^2 Var[X].$$

Moreover, let $\overline{X}$ be the sample mean. Xhen,

$$E[g(\overline{X})] \approx g(\mu)$$
$$Var[g(\overline{X})] \approx [g'(\mu)]^2 \frac{Var[X]}{k}.$$

Example 11.5.3. Let X and Y are random variables with means μ_X and μ_Y, respectively. Let $g(x, y) = x/y$. $\frac{\partial g}{\partial x} = \frac{1}{\mu_X}, \frac{\partial g}{\partial y} = -\frac{\mu_X}{\mu_Y^2}$.

We have

$$E[\frac{X}{Y}] \approx \frac{\mu_X}{\mu_Y}$$

and

$$E[\frac{X}{Y}] \approx \frac{1}{\mu_Y^2} Var[X] + \frac{\mu_X^2}{\mu_Y^4} Var[Y] - 2\frac{\mu_X}{\mu_Y^3} Cov(X, Y).$$

11.6 Moment generating functions and characteristic functions

11.6.1 Moment generating function

Moment generation functions are functions widely used to generate moments of random variables, and more importantly, characterize distributions.

Definition 11.6.1 (moment generating function (mgf)). *The moment generating function (mgf) of a random variable X is given as*

$$M_X(t) = E[e^{tX}],$$

provided that the expectation exists for t in some neighborhood of 0.

Remark 11.6.1 (existence of moment generating function). If the expectation does not exist for some t in the neighborhood of 0, then moment generating function does not exist.

Example 11.6.1. Let X be a random variable with normal distribution $N(0,1)$, then the moment generating function is

$$m_X(t) = E[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} = \exp(\frac{1}{2}t^2).$$

The most direct application of mgfs is to generate moments.

Lemma 11.6.1 (generating moments). *Let X be a random variable with moment generating function $M_X(t)$. Under the assumption of exchange expectation and differential is legitimate, for $n > 1$, then*

- $$E[X^n] = M_X^{(n)}(0) = \frac{d^n M_X(t)}{dt^n}|_{t=0}.$$
- $$M_X(t) = 1 + \sum_{n=1}^{\infty} \frac{M^{(n)}(0)}{n!} t^n = 1 + \sum_{n=1}^{\infty} \frac{E[X^n]}{n!} t^n.$$

Proof. (1)

$$M_X(t) = \int e^t x f(x) dx$$
$$M_X^{(n)}(t) = \int x^n e^t x f(x) dx$$
$$M_X^{(0)}(t) = \int x^n f(x) dx$$

(2) Use Taylor expansion. □

More important application of moment generating functions is to characterize distributions.

Theorem 11.6.1 (fundamental relationship between distribution and moment generating functions). *[7, p. 65] Let $F_X(x)$ and $F_Y(y)$ be two cdfs all of whose moments exist. We have*

- *If X and Y have bounded support, then $F_X(u) = F_Y(u)$ for all u if and only if*

$$E[X^r] = E[Y^r]$$

  for all integers $r = 0, 1, 2, ...$
- ***(uniqueness)**If the moment generating functions exist and $M_X(t) = M_Y(t)$ for all t in some neighborhood of o, then $F_X(u) = F_Y(u)$ for all u.*

One of most important applications of moment generation functions is to study the distribution after some transformations, such as addition and scaling. In the following, we will show the moment generation function of a random variable after such transformation.

Lemma 11.6.2. *Let $Y = g(X)$, where g is a monotone function, let $m(x)$ be a function, then*

$$\int_{\mathcal{Y}} m(y) f_Y(y) dy = \int_{\mathcal{X}} m(g(x)) f_X(x) dx$$

Proof. From the change of variable theorem [Lemma 11.4.5], we have

$$\int_{\mathcal{Y}} m(y) f_Y(y) dy = \int_{\mathcal{Y}} m(y) f_X(g^{-1}(y)) \left| dx/dy \right| dy$$

Let $y = g(x), dy = (dy/dx)dx$, then

$$\int_{\mathcal{Y}} m(y) f_X(g^{-1}(y)) \left| dx/dy \right| dy = \int_{\mathcal{X}} m(g(x)) f_X(x) dx.$$

□

Theorem 11.6.2 (addition and scaling property of mgf). *Let X and Y be two independent random variables, then*

- $M_{X+Y}(t) = M_X(t)M_Y(t)$
- *If $Z = aX + b$, then $M_Z(t) = e^{bt}M_X(at)$*

Proof. (1) Let $Z = X + Y$, $f_Z(z) = \int f_X(z-y)f_Y(y)dy$, then

$$M_Z = \int e^{zt} f_Z(z)dz = \int e^{zt} \int f_X(z-y)f_Y(y)dy = \int e^{(z-y)t} f_X(z-y)dz \int e^{yt} f_Y(y)dy$$

let $w = z - y$, then $dw = dz - dy, dzdy = dydw((dy)^2 = 0)$, we have

$$\int e^{(z-y)t} f_X(z-y)dz \int e^{y} f_Y(y)dy = \int e^{wt} f_X(w)dw \int e^{yt} f_Y(y)dy = M_X(t)M_Y(t)$$

(2) From Lemma 11.6.2, $M_Z(t) = \int e^{zt} f_Z(z)dz = \int e^{axt+bt} f_X(x)dx = e^{bt}M_X(at)$ □

Example 11.6.2. Let X be a random variable with normal distribution $N(0,1)$, then the moment generating function is

$$m_X(t) = \exp(\frac{1}{2}t^2).$$

If Y is a random variable with normal distribution $N(\mu, \sigma^2)$, then the moment generating function is

$$m_Y(t) = \exp(\mu t + \frac{1}{2}\sigma^2 t^2).$$

Moment generating functions can also be used to characterize the independence between two random variables.

Lemma 11.6.3 (independence from moment generating functions). *link Let X and Y be two random variables with space $\mathbb{R}$. Assume the mgfs for X, Y and $X + Y$ exist at the neighborhood of o. If for all t_X, t_Y in the neighborhood of o we have*

$$E[\exp(t_X X + t_Y Y)] = E[\exp(t_X X)]E[\exp(t_Y Y)],$$

then X and Y are independent.

Proof. Let (U, V) be such that U and V are independent; moreover, U and X have the same distribution and V and Y have the same distribution.

$$\begin{aligned} E[\exp(t_X X + t_Y Y)] &= E[\exp(t_X X)]E[\exp(t_Y Y)] \\ &= E[\exp(t_X U)]E[\exp(t_Y V)] = E[\exp(t_X U + t_Y V)], \end{aligned}$$

Therefore (X, Y) and (U, V) have the same joint distribution; that is, X and Y are independent. □

11.6.2 Characteristic function

A concept closely related to moment generating functions is the characteristic function. Similar to mgf, characteristic function is mainly used to characterize distributions. The characteristic function as the Fourier transform of the density function $f(x)$.

Definition 11.6.2 (characteristic function). *Given a random variable X with probability measure P, its characteristic function is given as*

$$\psi_X(t) = E[e^{itX}] = \int_{-\infty}^{\infty} e^{itx} dP(x) = \int_{-\infty}^{\infty} e^{itx} f(x) dx.$$

Because $\left|e^{itx} f(x)\right| \leq \left|f(x)\right|$ is L^1 integrable, then characteristic function always exists. Every distribution has a unique characteristic function; and to each characteristic function there corresponds a unique distribution of probability.

Remark 11.6.2 (Moment generating functions vs characteristic functions).

- Characteristic function always exists, whereas moment generating function not necessarily exists.
- Characteristic function is useful when we want to develop theory for more general pdf.

Lemma 11.6.4 (recovering probability distribution from characteristic function). *Let $\psi_X(t)$ be the characteristic function of random variable X. Then we can obtain its probability density function via*

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \psi_X(t) \exp(-itx) dt$$

Proof. Use the property of Fourier transform [Lemma 5.7.1]:

$$\begin{aligned}
\frac{1}{2\pi} \int_{-\infty}^{\infty} \psi_X(t) \exp(-itx) dt &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{itx} dP(x) \\
&= \int_{-\infty}^{\infty} e^{itx'} f(x') \exp(-itx) dt dx' \\
&= \int_{-\infty}^{\infty} f(x') \delta(x - x') dx' = f(x)
\end{aligned}$$

□

11.6.3 Joint moment generating functions for random vectors

We can extend moment generating functions to random vectors.

Definition 11.6.3 (joint moment generating function). *The joint moment generating function for a random vector $X = (X_1, ..., X_n)^T$ is defined as*

$$m_X(t) = E[\exp(t^T X)]$$

where $t \in \mathbb{R}^n, m_X(t) \in \mathbb{R}$, if the expectation exists in the neighborhood of the origin.

Lemma 11.6.5 (constructing joint moment generating function). *Let X be a K-dimensional random vector with a joint mgf $M_X(t)$, then we have*

- *If $X_1, X_2, ..., X_K$ are mutually independent of each other, then $M_X(t) = M_{X_1}(t_1)...M_{X_K}(t_K)$*
- *Let A be a matrix and b a vector, then $Z = AX + b$ has joint mgf given as*

$$M_Z(t) = e^{t^T b} M_X(A^T t)$$

Proof. Directly from definitions. □

Lemma 11.6.6 (cross moment generation). *Let X be a K dimensional random vector possessing a joint mgf $M_X(t)$, then*

$$\mu_X(n_1, n_2, ..., n_K) = E[X_1^{n_1} X_2^{n_2} ... X_K^{n_K}]$$

is given by

$$\mu_X(n_1, n_2, ..., n_K) = \frac{\partial^{n_1 + ... + n_K} M_X(t_1, ..., t_K)}{\partial t_1^{n_1} ... \partial t_K^{n_K}} |_{t=0}$$

Remark 11.6.3 (some applications). With joint mgf, we can evaluate the mean and covariance easily. For example, $E[X_1]$ can be obtained by setting $n_1 = 1, n_2 = 0, n_K = 0$. $E[X_i X_j]$ can be obtained by setting $n_i = n_j = 1$.

11.6.4 Cumulants

Definition 11.6.4 (cumulant-generating function, cumulant).

- *The **cumulant-generating function** $K(t)$ of a random variable X is defined by*

$$K(t) = \ln E[\exp(tX)] = \ln M_X(t),$$

where $M_X(t)$ is the moment generating function of X.

- *The **cumulants** κ_n are obtained via*

$$\kappa_n = K^{(n)}(0),$$

such that

$$K(t) = \sum_{n=1}^{\infty} \kappa_n \frac{t^n}{n!}.$$

Lemma 11.6.7 (connections between cumulants and moments). *Let $\mu_i, i = 1, 2, \ldots$ denote the central moments,i.e. $\mu_i = E[(X - E[X])^i]$ of a distribution of a random variable X.Let $m_i, i = 1, 2, \ldots$ denote the cumulants of the same distribution. Let $\kappa_i, i = 1, 2, \ldots$ denote the cumulants of the same distribution. Assume the existence of moment generating function. Then*

- $$\ln(1 + \sum_{n=1}^{\infty} \frac{m_n}{n!} t^n) = \sum_{n=1}^{\infty} \frac{\kappa_n}{n!} t^n$$

- *explicitly, we have*

$$\begin{aligned}
\kappa_1 &= m_1 \\
\kappa_2 &= m_2 - m_1^2 = \mu_2 \\
\kappa_3 &= \mu_3 \\
\kappa_4 &= \mu_4 - 3\mu_2^2 \\
\kappa_5 &= \mu_5 - 10\mu_3\mu_2.
\end{aligned}$$

Proof. (1) Based on the definition,

$$\sum_{n=1}^{\infty} \kappa_n \frac{t^n}{n!} = K(t) = \ln E[\exp(tX)] = \ln M_X(t) = \ln(1 + \sum_{n=1}^{\infty} \frac{m_n}{n!} t^n),$$

where we use the properties of moment generating functions [Lemma 11.6.1]. (2) Use Taylor expansion for $\ln(1+x)$ [Lemma 3.6.4] given by

$$\ln(1+x) = \sum_{n=1}^{\infty} (-1)^{n-1} \frac{x^n}{n} = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + ...$$

and then match the coefficients for t^n. □

Example 11.6.3. Consider a Gaussian distribution given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(x-\mu)^2}{2\sigma^2}).$$

Then

- the cumulant generating function is given by

$$K(t) = \ln(e^{\mu} e^{\sigma^2 \frac{t^2}{2}}) = \mu t + \frac{\sigma^2 t^2}{2}.$$

- the cumulants are given by

$$\kappa_1 = \mu, \kappa_2 = \sigma^2, \kappa_n = 0, \forall n > 2.$$

11.7 Conditional expectation

11.7.1 General intuitions

Consider a random variable defined on a probability space $(\Omega, \mathcal{F}, P)$ and a sub-σ-algebra $\mathcal{G}$ of $\mathcal{F}$ ($\mathcal{G}$ is a σ-algebra and $\mathcal{G} \subset \mathcal{F}$). We have the following situations:

1. If X is independent of $\mathcal{G}$, then the information in $\mathcal{G}$ provides no help in determining the value X. In this case, $E[X|\mathcal{G}] = E[X]$.
2. If X is $\mathcal{G}$ measurable, then the information in $\mathcal{G}$ can fully determine X. In this case, $E[X|\mathcal{G}] = X$.
3. In the intermediate case, we can use information in $\mathcal{G}$ to estimate but not precisely evaluate X. The *conditional expectation* of X given $\mathcal{G}$ is such an estimate.
4. If $\mathcal{G}$ is the trivial σ algebra $\{\emptyset, \Omega\}$, then $\mathcal{G}$ barely contains any information: $E[X|\mathcal{G}] = E[X]$.

Another understanding in terms of random variables are: $E[X|Y]$ is the function of Y that best approximates X. We consider a extreme case. Suppose that X is itself a function of Y, then the function of Y that best approximates X is X itself, i.e., $E[g(Y)|Y] = X = g(Y)$; If X is independent of Y, then the best estimate we can give is $E[X|Y] = E[X]$.

As a summary, we have

Definition 11.7.1 (conditional expectation as least-squared-best predictor). *[8] If $E[X^2] < \infty$, then the conditional expectation $Y = E[X|\mathcal{G}]$ is a version of the orthogonal projection of X onto the space $L^2(\Omega, \mathcal{G}, P)$. Hence, Y is the lease-squared-best $\mathcal{G}$-measurable predictor of X: among all $\mathcal{G}$-measurable functions, Y minimizes*

$$E[(Y - X)^2].$$

Remark 11.7.1. Note that the discussion on the existence and uniqueness of such Y can be found at [8][9, p. 28].

11.7.2 Formal definitions

Definition 11.7.2 (sub σ algebra). *Let X be a set and let $\mathcal{F}, \mathcal{G}$ be two σ algebras on X. then $\mathcal{G}$ is said to be sub-σ algebra of $\mathcal{F}$ if $\mathcal{G} \subseteq \mathcal{F}$.*

Definition 11.7.3 (conditional expectation as a random variable). *[4, p. 68] Let $(\Omega, \mathcal{F}, P)$ be a probability space, let $\mathcal{G}$ be a **sub-σ algebra** of $\mathcal{F}$, and let X be a random variable that is either non-negative or integrable. The conditional expectation of X given $\mathcal{G}$, denoted $E[X|\mathcal{G}]$ is a **random variable** that satisfies:*

1. *(**measurability**) $E[X|\mathcal{G}]$ is $\mathcal{G}$ measurable*
2. *(partial averaging): For any element A in $\mathcal{G}$,*

$$\int_A E[X|\mathcal{G}](\omega)dP(\omega) = \int_A X(\omega)dP(\omega).$$

In particular,

- *if $\mathcal{G} = \mathcal{F}$, then $E[X|\mathcal{G}] = X$.*
- *If $G = \{\emptyset, \Omega\}$, then $E[X|\mathcal{G}] = E[X]$.*

[a]

[a] The meaning of X is $\mathcal{G}$ measurable can be understood as $\sigma(X) \subseteq \mathcal{G}$.

Remark 11.7.2.

- the filtration $\mathcal{G}$ in $E[X|\mathcal{G}]$ has to be $\mathcal{G} \subseteq \mathcal{F}$, otherwise P is defined for some elements in cG.
- the Partial averaging property reflects the **consistence** requirement between the new random variable $E[X|\mathcal{G}]$ and the old random variable X.
- If $\mathcal{G}$ is the σ algebra generated by some other random variable W, then we generally write $E[X|W]$ instead of $E[X|\sigma(W)]$.
- if $G = \{\emptyset, \Omega\}$, then the only $\mathcal{G}$- measurable function is a constant function. Among all the constant functions, the function that satisfies the partial averaging property is the expectation.

Note 11.7.1 (interpreting partial averaging property in partition set)**.** Consider the case where $\mathcal{G}$ is countable. Let $\mathcal{P}$ be the smallest partition set of $\mathcal{G}$. Then the random variable $E[X|\mathcal{G}]$ can only take countable many values. In particular, the partial averaging property implies

$$E[X|\mathcal{G}](A_i) = \int_A X(\omega)dP(\omega) \forall A_i \in \mathcal{P}.$$

That is, $E[X|\mathcal{G}]$ can be viewed as a mapping from Ω to $\mathbb{R}$ that has been **coarsened via local averaging**.

Note 11.7.2 (Generalization on expectation). When we talk about expectation, there are two items we should consider: which measure the expectation is taken with respect to and which filtration the expectation is taken with respect to.

- We can view expectation as a special case of conditional expectation: for example
$$E[X] = E[X|\mathcal{G}], \mathcal{G} = \{\emptyset, \Omega\}.$$
- Conditional expectations with respect to different measure can equal if the two measures agree on the filtration. For example,
$$E_P[X|\mathcal{G}] = E_Q[X|\mathcal{G}],$$
if $P(A) = Q(A), \forall A \in \mathcal{G}$.

11.7.3 Different versions of conditional expectation

Remark 11.7.3. For different versions of conditional expectation, see [9, p. 17] for details.

11.7.3.1 *Conditioning on an event*

Definition 11.7.4. *For any integrable random variable η and any event $B \in \mathcal{F}$ such that $P(B) \neq 0$, the conditional expectation given B is defined as*
$$E[\eta|B] = \frac{\int_B \eta dP}{\int_B dP} = \frac{1}{P(B)} \int_B \eta dP$$

11.7.3.2 *Conditioning on a discrete random variable as a new random variable*

Definition 11.7.5. *Let X be an integrable random variable, let Y be a discrete random variable. Then the conditioning expectation of X given Y is defined to be a random variable $E[X|Y]$ such that*
$$E[X|Y](\omega) = E[X|\{Y(\omega) = y_i\}]$$

Lemma 11.7.1. *If X is an integrable random variable, and Y is a discrete random variable, then*

- *$E[X|Y]$ is $\sigma(Y)$-measurable*

- *For any $A \in \sigma(Y)$:*

$$\int_A E[X|Y]dP = \int_A XdP$$

Proof. When Y is a discrete random variable, $E[X|Y]$ can only take discrete values. For any Borel set on $\mathbb{R}$, we find the inverse image $B \in \sigma(Y)$. Therefore it is measurable. (2) directly form partial averging property of conditional expectation. □

11.7.3.3 *Condition on random variable vs. event vs σ algebra*

- Conditional expectations for discrete random variables, such as $E[X|Y = 2], E[X|Y = 5]$ are numbers. These are examples of condition on events. $E[X|Y]$ can be interpreted as $E[X|Y = y]$, a function depends on y.
- When we write $E[X|Y]$, we should interpret as conditioning on the σ algebra generated by Y.

11.7.4 Properties

11.7.4.1 *Linearity*

Lemma 11.7.2 (linearity of conditional expectation). *[4, p. 69] Let $(\Omega, \mathcal{F}, P)$ be a probability space, let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$, and let X, Y be integrable random variables. We have:*

$$E[c_1X + c_2Y|\mathcal{G}] = c_1E[X|\mathcal{G}] + c_2E[Y|\mathcal{G}].$$

Similarly, let Z be a random variable. We have

$$E[c_1X + c_2Y|Z] = c_1E[X|Z] + c_2E[Y|Z].$$

Proof. (1) First, $c_1E[X|\mathcal{G}]$ is $\mathcal{G}$ measurable, $c_2E[Y|\mathcal{G}]$ is $\mathcal{G}$ measurable, therefore, $E[c_1X + c_2Y|\mathcal{G}] = c_1E[X|\mathcal{G}] + c_2E[Y|\mathcal{G}]$ is $\mathcal{G}$ measurable [Definition 11.5.1]. (2) For every $A \in \mathcal{G}$,

$$\begin{aligned}
&\int_A (c_1E[X|\mathcal{G}](\omega) + c_2E[Y|\mathcal{G}](\omega))dP(\omega) \\
=&\int_A (c_1E[X|\mathcal{G}](\omega) + c_2E[Y|\mathcal{G}](\omega))dP(\omega) \\
=&c_1\int_A (E[X|\mathcal{G}](\omega)dP(\omega) + c_2\int_A E[Y|\mathcal{G}](\omega))dP(\omega) \\
=&c_1\int_A X(\omega)dP(\omega) + c_2\int_A Y(\omega)dP(\omega) \\
=&\int_A c_1X(\omega) + c_2Y(\omega)dP(\omega)
\end{aligned}$$

that is $E[c_1X + c_2Y|\mathcal{G}]$ satisfies the partial averaging property. □

11.7.4.2 *Taking out what is known*

Lemma 11.7.3. *Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$. Let X, Y be integrable random variables. If XY is integrable, X is $\mathcal{G}$-measurable, then*

- $$E[XY|\mathcal{G}] = XE[Y|\mathcal{G}], E[g(X)Y|\mathcal{G}] = g(X)E[Y|\mathcal{G}].$$
- $$E[X|\mathcal{G}] = X, E[X|X] = X, E[g(X)|X] = g(X)$$

Proof. Note that from Definition 11.5.1, $g(X), XY, g(X)Y$ are all $\mathcal{F}$ measurable random variables. □

11.7.4.3 *Law of total expectation*

In Theorem 11.2.1, we discuss the law of total probability,

$$P(C) = P(C \cap C_1) + P(C \cap C_2)... + P(C \cap C_k) = \sum_{i=1}^{k} P(C_i)P(C|C_i),$$

where $C_1, C_2, ..., C_k$ are mutual disjoint subsets that partition the sample space Ω. The generalization from probability to expectation is the following law of total expectation.

Theorem 11.7.1 (law of total expectation). *Let X be a random variable, Let $A_1, ..., A_n \in \mathcal{F}$ be the partition of the sample space, then*

$$E[X] = \sum_{i=1}^{n} E[X|A_i]P(A_i)$$

In concise form, we have

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

where Y is the random variable defined on measure space $(\Omega, \sigma(A_1, ..., A_n))$.

Proof. We consider the special cases where X, Y are discrete random variable taking values in $\mathcal{X}$ and cY.

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

Assume the series is finite so that we can exchange the summations, we have

$$\begin{aligned} \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} x \cdot P(X = x, Y = y) &= \sum_{x\in\mathcal{X}} x \sum_{y\in\mathcal{Y}} P(X = x, Y = y) \\ &= \sum_{x\in\mathcal{X}} x \cdot P(X = x) \\ &= E[X]. \end{aligned}$$

□

Example 11.7.1. Suppose we want to estimate the mean height of people in a certain region. Suppose we already have following estimates:

- the estimate of mean height of male and female, respectively, in this region;
- the estimate of male and female populations in the region.

We can use the law of total expectation to estimate the mean height of the total population in the following way:

Let H be the height of a randomly sampled person, and let M be the event that the person is male and F the event that the person is female. Then the mean height $E[H]$ can be computed via

$$E[H] = E[H|M]P(M) + E[H|F]P(F).$$

11.7.4.4 *Law of iterated expectations*

Lemma 11.7.4 (iterated conditioning). *If $\mathcal{H}, \mathcal{G}$ are both σ algebra on Ω, and $\mathcal{G} \subset \mathcal{H}$ (in some sense $\mathcal{G}$ has less information), then for random variable X, we have*

$$E[E[X|\mathcal{H}]|\mathcal{G}] = E[X|\mathcal{G}]$$

$$E[E[X|\mathcal{G}]|\mathcal{H}] = E[X|\mathcal{G}].$$

In particular,

$$E[E[X|\mathcal{G}]] = E[X],$$

or equivalently, in terms of conditioning on random variables, we have

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

Proof. (1)(a) First $E[X|\mathcal{G}]$ is $\mathcal{G}$-measurable. (b) For any $A \in \mathcal{G} \subseteq H$, we have

$$\begin{aligned}
&\int_A E[E[X|\mathcal{H}]|\mathcal{G}](\omega)dP(\omega) \\
=&\int_A E[X|\mathcal{H}](\omega)dP(\omega) \\
=&\int_A X(\omega)dP(\omega) \\
=&\int_A E[X|\mathcal{G}](\omega)dP(\omega)
\end{aligned}$$

(2) Note that the random variable $E[X|\mathcal{G}]$ is $\mathcal{G}$-measurable therefore $\mathcal{H}$-measurable.

□

11.7.4.5 *Conditioning on independent random variable/σ algebra*

Lemma 11.7.5. *[4, p. 70]Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$. Let X, Y be integrable random variables. Let f be Borel measurable function and $f(X)$ be integrable.*

- *If $\sigma(X)$ and $\mathcal{G}$ are independent, then*

$$E[X|\mathcal{G}] = E[X], E[g(X)|\mathcal{G}] = E[g(X)].$$

- *If X and Y are independent, then*

$$E[X|Y] = E[X|\sigma(Y)] = E[X].$$

Proof. (1)(a) $E[X]$ is a constant, therefore is $\mathcal{G}$ measurable. (b)(informal) Consider the special case where $X = \mathbf{1}_B$, where $B \in \mathcal{F}$ but B is independent of $\mathcal{G}$. Then

$$\int_A X(\omega)dP(\omega) = P(A \cap B) = P(A)P(B) = E[X]P(A) = E[X]\int_A dP(\omega) = \int_A E[X]dP(\omega).$$

Since X can be represented by the sum of indicator function, such relation can hold when X is an arbitrary random variable.(See reference for more details). □

11.7.4.6 *Least Square minimizing property*

Lemma 11.7.6 (least square minimizing property of conditional expectation). *Let $Y \in \mathcal{L}_2(\Omega, \mathcal{G}, P)$ and $\mathcal{F}$ be a sub-σ of $\mathcal{G}$, then*

$$E[(Y - E[Y|\mathcal{F}])^2] = \min\{E[(Y - Z)^2], \forall Z \in \mathcal{L}_2(\Omega, \mathcal{F}, P)\}$$

Proof. For any $Z \in \mathcal{L}_2(\Omega, \mathcal{F}, P)$, we have

$$\begin{aligned}
&E[(Y - E[Y|\mathcal{F}])^2] \\
&= E[(Y - Z + Z - E[Y|\mathcal{F}])^2] \\
&= E[(Y - Z)^2] + E[(Z - E[Y|\mathcal{F}])^2] + 2E[(Y - Z)(Z - E[Y|\mathcal{F}])] \\
&= E[(Y - Z)^2] + E[(Z - E[Y|\mathcal{F}])^2] + 2E[E[(Y - Z)(Z - E[Y|\mathcal{F}])|\mathcal{F}]] \\
&= E[(Y - Z)^2] + E[(Z - E[Y|\mathcal{F}])^2] + 2E[E[(Y - Z)|\mathcal{F}](Z - E[Y|\mathcal{F}])] \\
&= E[(Y - Z)^2] + E[(Z - E[Y|\mathcal{F}])^2] - 2E[(Z - E[Y|\mathcal{F}])(Z - E[Y|\mathcal{F}])] \\
&= E[(Y - Z)^2] - E[(Z - E[Y|\mathcal{F}])^2] \leq E[(Y - Z)^2]
\end{aligned}$$

Note that we use $E[(Y - Z)(Z - E[Y|\mathcal{F}])|\mathcal{F}] = (Z - E[Y|\mathcal{F}])E[(Y - Z)|\mathcal{F}]$ since $(Z - E[Y|\mathcal{F}])$ is $\mathcal{F}$ measurable. □

11.8 The Hilbert space of random variables

11.8.1 Definitions

Definition 11.8.1. *The vector space $L^2(\Omega, \mathcal{F}, P)$ of real-valued random variables on $(\Omega, \mathcal{F}, P)$ can be defined as the Hilbert space of random variables with finite second moment. The inner product is then defined as*

$$\langle x, y \rangle = E[xy].$$

The norm of a random variable is

$$\|X\| = \sqrt{E[X^2]}.$$

Lemma 11.8.1 (correlation and orthogonality for zero mean random variables). *Let X and Y be two zero mean random variables in the Hilbert space $L^2(\Omega, \mathcal{F}, P)$. Then X and Y are uncorrelated if and only if they are orthogonal, i.e., $\langle X, Y \rangle = 0$.*

Proof. (1) If $\langle X, Y \rangle = 0$, then

$$E[XY] = E[X]E[Y] + Cov(X, Y) = 0 \implies Cov(X, Y) = 0$$

. (2) If $Cov(X, Y) = 0$, then

$$\langle X, Y \rangle = E[XY] = E[X]E[Y] + Cov(X, Y) = 0.$$

□

11.8.2 Subspaces, projections, and approximations

Theorem 11.8.1 (projection onto closed subspace , recap). *Let U be a closed subspace of L^2 and $X \in L^2$. Then the projection of X onto U is the vector/random variable $V \in U$ such that*

- $$\langle X - V, u \rangle = E[(X - V)u] = 0, \forall u \in U$$
- *V is unique;*
- *V is minimizer,i.e., $\|X - V\|^2 \leq \|X - u\|^2, \forall u \in U$.*

[a].

[a] Note that in a Hilbert space(also a normed linear space), any finite-dimensional subspace is closed [Theorem 5.2.1]

Proof. See the projection theorem [Theorem 5.3.5] guarantees the existence of solution. □

Lemma 11.8.2 (projection onto the subspace of constant random variables).

- *Let real-valued random variable $X \in L^2$, we define the root mean square error function by*
$$d_2(X,t) = \|X - t\|_2 = \sqrt{E[(X-t)^2]}, t \in \mathbb{R}$$
then $d_2(X,t)$ is minimized when $t = E[X]$ and that the minimum value is $\sqrt{Var[X]}$.
- *Let real-valued random variable $X \in L^2$, we define the 1d subspace $W = \{a : a \in \mathbb{R}\}$(the subspace spanned by constant random variable 1). Then the projection of X onto W is $E[X]$.*

Proof. (1)directly minimize with respect t. (2) We can see that the orthogonality condition implies that
$$0 = \langle X - a, b \rangle = E[(X-a)b] = 0, \forall b \in \mathbb{R},$$
which gives $a = E[X]$. □

Theorem 11.8.2 (best linear predictor for random variables).

- *Given $X, Y \in L^2$, the best linear predictor for Y given X is to find a projection onto the subspace $W = \{a + bX : a \in \mathbb{R}, b \in \mathbb{R}\}$(the subspace spanned by random variable 1 and X), given as*
$$L(Y|X) = E[Y] + \frac{Cov(X,Y)}{Var(X)}(X - E[X])$$
and the variance/mean square error for the prediction is
$$Var(Y - L(Y|X)) = Var(Y) - \frac{Cov(X,Y)^2}{Var(X)}.$$
- *Given $X_1, X_2, ..., X_n, Y \in L^2$, the best linear predictor for Y given $X_1, X_2, ..., X_n$ is*
$$L(Y|X) = E[Y] + \sum_{i=1}^{n}(X_i - E[X_i])[\sum_{j=1}^{n}(\Sigma_{XX}^{-1})_{ij}Cov(X_j, Y)],;$$

or in vector form

$$L(Y|X) = E[Y] + (X - E[X])^T \beta,$$

where $X = (X_1, X_2, ..., X_n)^T, \beta = \Sigma_{XX}^{-1}\Sigma_{XY}$. *In particular, if* $Cov(X_i X_j) = Var[X_i]\delta_{ij}$, *then*

$$L(Y|X) = E[Y] + \sum_{i=1}^{n} \frac{Cov(X_i, Y)}{Var(X_i)}(X_i - E[X_i]).$$

- *The estimation error is given by*

$$E[(Y - L(Y|X))^2] = Var[Y] - \Sigma_{XY}^T(\Sigma_{XX}^{-1})\Sigma_{XY}.$$

- *The single coefficient associated with* X_i *is given by*

$$\beta_i = \frac{Cov(Y - L(Y|X_{-i}), X_i - L(X_i|X_{-i})}{Var[X_i - L(X_i|X_{-i})]},$$

where X_{-i} *denotes the subspace associated with span* $\{1, X_1, ..., X_{i-1}, X_{i+1}, ..., X_n\}$.
- *un-correlation of the residual and* X*:*

$$Cov(Y - L(Y|X), X) = 0.$$

Proof. (1) To verify that $L(Y|X)$ is the projection, we only need to verify the orthogonality conditions [Theorem 5.3.5]:

$$\langle Y - L(Y|X), X \rangle = 0, \langle Y - L(Y|X), 1 \rangle = 0.$$

We have

$$\begin{aligned}
\langle Y - L(Y|X), X \rangle &= E[(Y - L(Y|X))X] \\
&= E[(Y - E[Y] - \frac{Cov(X, Y)}{Var(X)}(X - E[X]))X] \\
&= E[(Y - E[Y])X] - E[\frac{Cov(X, Y)}{Var(X)}(X - E[X]))X] \\
&= Cov(X, Y) - Cov(X, Y) \\
&= 0
\end{aligned}$$

where we used the fact that $E[X(X - E[X])] = Var[X]$. For another,

$$\begin{aligned}\langle Y - L(Y|X), 1 \rangle &= E[(Y - L(Y|X))] \\ &= E[(Y - E[Y] - \frac{Cov(X,Y)}{Var(X)}(X - E[X]))] \\ &= 0 - 0 \\ &= 0.\end{aligned}$$

The variance is given by

$$\begin{aligned}&Var[Y - L(Y|X)] \\ &= E[(Y - L(Y|X))^2] \\ &= E[(Y - E[Y] - \frac{Cov(X,Y)}{Var(X)}(X - E[X]))(Y - E[Y] - \frac{Cov(X,Y)}{Var(X)}(X - E[X]))] \\ &= E[(Y - E[Y])^2] + \frac{Cov(X,Y)^2}{Var(X)} - 2E[(Y - EY)(\frac{Cov(X,Y)}{Var(X)}(X - E[X]))] \\ &= E[(Y - E[Y])^2] + \frac{Cov(X,Y)^2}{Var(X)} - 2\frac{Cov(X,Y)^2}{Var(X)} \\ &= E[(Y - E[Y])^2] - \frac{Cov(X,Y)^2}{Var(X)}\end{aligned}$$

(2)

We can obtain the vector form via the optimization

$$\min f = E[(Y - \beta_0 - \beta^T X)^2]$$

over β_0, β_1, we have

$$f(\beta_0, \beta_1) = E[(Y^2 + \beta_0^2 + (\beta^T X)^2 + 2\beta_0\beta^T X - 2\beta_0 Y - 2Y\beta^T X)]$$

The first order condition on β_0 gives that

$$\beta_0 = E[Y] - \beta^T E[X];$$

Plug in β_0 and the first order condition on β_1 gives that

$$\begin{aligned}&f(\beta_0, \beta_1) = E[((Y - EY)^2 - 2\beta^T(X - EX)(Y - EY) + \beta^T E[(X - EX)(X - EX)^T]\beta \\ \implies &\partial f/\partial\beta = -2E[(X - EX)(Y - EY)] + 2E[(X - EX)(X - EX)^T]\beta = 0 \\ \implies &\beta = (E[(X - EX)(X - EX)^T])^{-1}E[(X - EX)(Y - EY)] = (\Sigma_{XX}^{-1})\Sigma_{XY}\end{aligned}$$

Note that the problem has semi-positive definite Hessian we are sure that the minimizer exists.

From the Hilbert space projection point of view, we can also verify the orthogonality conditions [Theorem 5.3.5]:

$$\langle Y - L(Y|X), X_k \rangle = 0, k = 1, 2, ..., n.$$

We have

$$\begin{aligned}
\langle Y - L(Y|X), X_k \rangle &= E[(Y - E[Y] + \sum_{i=1}^{n}(X_i - E[X_i])[\sum_{j=1}^{n}(\Sigma_{XX}^{-1})_{ij}Cov(X_j, Y)]))X_k] \\
&= E[(Y - E[Y])X_k] - E[(\sum_{i=1}^{n}(X_i - E[X_i])[\sum_{j=1}^{n}(\Sigma_{XX}^{-1})_{ij}Cov(X_j, Y)])X_k] \\
&= Cov(X_k, Y) - Cov(X_j, Y)\delta_{jk} \\
&= 0
\end{aligned}$$

where we used the fact that

$$\sum_{i=1}^{n}(X_i - E[X_i]) \sum_{j=1}^{n}(\Sigma_{XX}^{-1})_{ij}X_k = \delta_{jk}.$$

Note that for an invertible matrix A, $\sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}A_{jk}^{-1} = \delta_{ik}$.

(3) Note that $L(Y|X)$ is unbiased because of the orthogonality condition

$$\langle Y - L(Y|X), 1 \rangle = 0 \implies E[(Y - L(Y|X))1] = E[Y] - E[L(Y|X)] = 0.$$

In the following we use the notation

$$\hat{Y} = L(Y|X), E[Y] = \mu_Y = E[L(Y|X)] = \mu_{\hat{Y}}.$$

We have

$$\begin{aligned}
E[(Y - \hat{Y})^2] &= E[((Y - \mu_Y) - (\hat{Y} - \mu_{\hat{Y}}))^2] \\
&= E[((Y - \mu_Y))^2] - 2E[(Y - \mu_Y)(\hat{Y} - \mu_{\hat{Y}})] + E[((\hat{Y} - \mu_{\hat{Y}}))^2] \\
&= Var[Y] - 2E[(Y - \mu_Y)(\hat{Y} - \mu_{\hat{Y}})] + Var[\hat{Y}] \\
&= Var[Y] - 2E[(Y - \mu_Y)(\beta^T X - \mu_{\hat{Y}})] + Var[\hat{Y}] \\
&= Var[Y] - 2Cov(Y, \beta^T X) + Var[\beta^T X] \\
&= Var[Y] - 2\beta^T Cov(Y, X) + \beta^T Cov(X, X)\beta \\
&= Var[Y] - 2\Sigma_{XY}^T(\Sigma_{XX}^{-1})Cov(Y, X) + \Sigma_{XY}^T(\Sigma_{XX}^{-1})Cov(X, X)(\Sigma_{XX}^{-1})\Sigma_{XY} \\
&= Var[Y] - \Sigma_{XY}^T(\Sigma_{XX}^{-1})\Sigma_{XY}
\end{aligned}$$

(4) Direct generalization from Hilbert space approximation theory [Theorem 5.4.4].

(5)

$$\begin{aligned} E[Y - L(Y|X), X - E[X]] &= E[Y - E[Y] - (X - E[X])^T\beta, X - E[X]] \\ &= Cov(X, Y) - Var[X]\beta \\ &= Cov(X, Y) - Cov(X, Y) \\ &= 0. \end{aligned}$$

□

Remark 11.8.1.

- The more correlated X and Y are, the more information X can provide to predict Y
- The more volatile X is, the less information X can provide.
- The magnitude of $\frac{Cov(X,Y)^2}{Var(X)}$ reflects the importance of X in prediction.

11.8.3 Connection to conditional expectation

Theorem 11.8.3 (conditional expectation with respect to a σ algebra as a projection).

- *Let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$, then the set*

$$U = \{X \in L^2 | X \text{ is measurable with respect to } \mathcal{G}\}$$

  is a subspace of L^2.
- *If $X \in L^2$, then $E[X|\mathcal{G}]$ is the projection of X onto the subspace U defined as*

$$U = \{X \in L^2 | X \text{ is measurable with respect to } \mathcal{G}\}.$$

Proof. (1) the zero element o is both $\mathcal{F}$ and $\mathcal{G}$ measurable. (2) If X, Y are $\mathcal{G}$ measurable, then $cX, X + Y$ are $\mathcal{G}$ measurable [Lemma 3.8.4]. (2) directly from the definition of conditional expectation [Definition 11.7.3]. □

Definition 11.8.2 (conditional expectation and projection). *The conditional expectation of $X \in L^2$ given $X_1, X_2, ..., X_n \in L^2$ is defined to be the projection of X onto*

*the closed subspace $M(X_1, X_2, ..., X_n)$ spanned by **all random variables of the form** $g(X_1, X_2, ..., X_n)$, where g is some measurable function $g : \mathbb{R}^n \to \mathbb{R}$, i.e.*

$$E[X|X_1, X_2, ..., X_n] = P_S[M(X_1, X_2, ..., X_n)].$$

Definition 11.8.3 (conditional expectation and projection onto a subspace, special case). *The conditional expectation of $X \in L^2$ given a closed subspace $S \subseteq L^2$, which contains the constant random variable 1, is defined to be the projection of X onto S,i.e.,*

$$E[X|S] = P_S[X].$$

Remark 11.8.2. Note that the subspace has to contains the constant random variable to make the definition and conditional expectation and projection match.

Remark 11.8.3.

- $$span(1, X_1, ..., X_n) \subseteq M(X_1, X_2, ..., X_n),$$
  therefore
  $$\left\|X - E[X|X_1, X_2, ..., X_n]\right\|^2 \leq \left\|X - E[X|span(1, X_1, X_2, ..., X_n)]\right\|^2.$$
- The definition of
  $$E[X|X_1, X_2, ..., X_n] = P_S[M(X_1, X_2, ..., X_n)]$$
  coincides with the usual definition of conditional expectation with respect to a σ algebra [Definition 11.7.3].
- The conditional expectation with respect to a subspace is not the general definition of conditional expectation.

Lemma 11.8.3 (conditional expectation and best predictor for multivariate normal random variables). *Let $(Y, X_1, X_2, ..., X_n), X = (X_1, X_2, ..., X_n)$ be a random vector with multivariate normal distribution with parameter*

$$\mu = [\mu_Y^T, \mu_X^T]^T, \Sigma = \begin{pmatrix} \Sigma_{YY} & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_{XX} \end{pmatrix}$$

Then

- *$Y|X_1, X_2, ..., X_n$ has the same distribution of*
  $$\hat{Y} = \mu_Y + \Sigma_{YX}\Sigma_{XX}^{-1}(X - \mu_X) + \epsilon,$$

conditioning on X and $\epsilon \sim N(0, \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY})$.

- $$E[Y|X_1, X_2, ..., X_n] = E[Y|span(1, X_1, X_2, ..., X_n)] = P_{span(1,X_1,X_2,...,X_n)}[Y].$$

  That is, the best predicator(in terms of minimum variance) given $X_1, X_2, ..., X_n$ is the best linear predictor.

Proof. From Theorem 14.1.2, the martinal distribution is Gaussian given by

$$N(\mu_Y + \Sigma_{YX}\Sigma_{XX}^{-1}(X - \mu_X), \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}).$$

Therefore, Y has the conditional expectation of

$$\mu_Y + \Sigma_{YX}\Sigma_{XX}^{-1}(X - \mu_X),$$

and the conditional variance of

$$\Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}.$$

□

11.9 Probability inequalities

11.9.1 Chebychev inequalities

Theorem 11.9.1 (General Chebychev's inequality). *Let X be a random variable and $g(x) \geq 0$. Then for any $r > 0$, we have:*

$$P(g(X) > r) \leq \frac{E[g(X)]}{r}.$$

Proof.

$$\begin{aligned} E[g(x)] &= \int_{\infty}^{\infty} g(x) f_X(x) dx \\ &\geq \int_{x:g(x)\geq r} g(x) f_X(x) dx \\ &\geq r \int_{x:g(x)\geq r} f_X(x) dx \\ &= rP(g(X) \geq r) \end{aligned}$$

□

Corollary 11.9.1.1 (Chebychev's inequality).

$$P(\frac{(X - EX)^2}{\sigma^2} \geq t^2) \leq \frac{1}{t^2} E[\frac{(X-\mu)^2}{\sigma^2}].$$

Or equivalently,

$$P(|X - EX| \geq t) \leq \frac{1}{t^2} Var[X], t \geq 0.$$

Proof. Let $g(X) = (X - \mu)^2/\sigma^2$ and use above theorem. □

Example 11.9.1. Let X be a random variable with mean $\mu = 4$ and standard deviation $\sigma = 1$. Then the probability that $X < 1$ or $X > 7$ is bounded by

$$P(|X - 4| > 3) \leq \frac{1^2}{3^2} = \frac{1}{9}.$$

Corollary 11.9.1.2. *Let $g : [0, \infty) \to [0, \infty)$ be a strictly increasing non-negative function, and set $h(x) = g(|x|)$ to obtain*

$$P(|X| \geq a) \leq \frac{E[g(|X|)]}{g(a)}$$

where $a > 0$.

Specifically, if $X \geq 0$, then we have the Markov's inequality given by

$$P(X \geq r) \leq E[X]/r.$$

Proof. Let $g(X) = X$ in the general Chebychev's inequality. □

11.9.2 Jensen's inequality

Lemma 11.9.1 (Jensen's inequality). *For any random variable X, if $g(x)$ is a convex function then*

$$E[g(X)] \geq g(E[X]).$$

Conversely, if $g(x)$ is a concave function, then

$$E[g(X)] \leq g(E[X]).$$

Proof. Here we will show the case of discrete random variables. Note that for convex function

$$g(\sum_{i=1}^{n} w_i x_i) \leq \sum_{i=1}^{n} w_i g(x_i), \forall w_i \geq 0, \sum_{i=1}^{n} w_i = 1, i = 1, ..., n.$$

□

Example 11.9.2 (Jensen's inequality applications).

- Use Jensen's inequality, it can be showed that $Var[X] \geq 0$. This is because $Var[X] = E[X^2] - E[X]^2$ and

  $$E[X]^2 \leq E[X^2]$$

  with $g(x) = x^2$.

- If a random variable X only has positive support, then we can show that

$$E[\frac{1}{X}] = \frac{1}{E[X]}.$$

  Here we use $g(x) = 1/x$, which is a convex function for $x > 0$.
- If a random variable X only has positive support, then we can show that

$$E[\log(X)] \leq \log(E[X]).$$

  Here we use $g(x) = \log(x)$, which is a concave function for $x > 0$.

11.9.3 Holder's, Minkowski, and Cauchy-Schwarz inequalities

Theorem 11.9.2 (Holder's inequality). *[10, p. 319] If $p, q > 1$ and $1/p + 1/q = 1$, then*

$$E[|XY|] \leq (E[|X|^p])^{1/p}(E[|Y|^q])^{1/q}.$$

The equality holds when there exists real numbers $\alpha, \beta > 0$ such that $\alpha|X|^p = \beta|Y|^q$ almost everywhere.

Proof. Let $A = (\int |x|^p \, dP)^{1/p} = E[|X^p|])^{1/p}$ and $B = (\int |y|^q \, dP)^{1/q} = (E[|Y^q|])^{1/q}$. Then let $a = |X| / A$, $b = |Y| / B$, and then apply Young's inequality:

$$ab = |XY| / AB \leq \frac{|X|^p}{pA^p} + \frac{|Y|^q}{qA^q} = \frac{a^p}{p} + \frac{b^q}{q}.$$

Integrate(Lebesgue) both sides use probability measure and notice that A, B are constant, $A^p = E[|X^p|]$, then

$$\frac{E[|XY|]}{(E[|X^p|])^{1/p}(E[|Y^q|])^{1/q}} \leq 1/p + 1/q = 1.$$

□

Remark 11.9.1. Let $q = p = 2$, and we get the Cauchy-Schwarz inequality.

Theorem 11.9.3 (Minkowski's inequality). *[10, p. 319] If $p \geq 1$, then*

$$(E[|X + Y|^p])^{1/p} \leq (E[|X|^p])^{1/p} + (E[|Y|^p])^{1/p}$$

Proof. Because L^p space are normed vector space, we can prove this using triangle inequality. □

Theorem 11.9.4 (Cauchy-Schwarz inequality). *[2][7, p. 187][11]*

- *Let X and Y be random variables with $E[X^2] < \infty, E[Y^2] < \infty$. Then*

$$\left|E[XY]\right| \leq \sqrt{E[X^2]E[Y^2]}.$$

The equality holds when there exist real numbers $\alpha, \beta > 0$ such that $\alpha|X|^2 = \beta|Y|^2$ almost everywhere.
Further more,

$$(Cov(X,Y))^2 \leq Var[X] \cdot Var[Y].$$

- *Let X and Y be two p dimensional random vectors with bounded variance. Then*

$$Var[Y] \geq Cov(Y,X)Var[X]^{-1}Cov(X,Y).$$

Proof. (1)(a) define inner product between two random variable as $\langle X, Y \rangle = \int xy\rho(x,y)dxdy$, since each random variable can be viewed as a functional. (b) Similarly we can use Holder's inequality. [Theorem 11.9.2]

A simple derivation: Since the covariance matrix of random vector (X, Y) much be positive semi-definite, we have

$$\left|Cov([XY])\right| = \begin{vmatrix} Var[X] & Cov(X,Y) \\ Cov(X,Y) & Var[Y] \end{vmatrix} \geq 0.$$

Expand the determinant, we have

$$Cov(X,Y)^2 \leq Var[X] \cdot Var[Y].$$

(2) See reference. □

Corollary 11.9.4.1 (bounds on correlations).

- *Let X and Y be two random variables with mean μ_X and μ_Y. Define correlation by*

$$\rho \triangleq \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sqrt{E[(X-\mu_X)^2]}\sqrt{E[(Y-\mu_Y)^2]}}.$$

Then

$$\left|\rho\right| \leq 1$$

.

- *Let $X_1, X_2, ..., X_n$ be the iid random sample of X. Let $Y_1, Y_2, ..., Y_n$ be the iid random sample of Y. Define sample correlation by*

$$\hat{\rho} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y - \overline{Y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}.$$

For each realizations of $X_1, ..., X_n, Y_1, ..., Y_n$, we have

$$|\hat{\rho}| \leq 1.$$

Proof. (1) From Cauchy-Schwartz inequality, we have

$$|\rho| = \frac{|E[(X - \mu_X)(Y - \mu_Y)]|}{\sqrt{E[(X - \mu_X)^2]}\sqrt{E[(Y - \mu_Y)^2]}} \leq 1.$$

(2) Suppose we have a realization of $x_1, x_2, ..., x_n, y_1, y_2, ..., y_n$, we can define a random variable X with probability $1/n$ in taking discrete values $x_1, x_2, ..., x_n$;similarly define a random variable Y. Then

$$|\hat{\rho}| = |\rho_{XY}| \leq 1.$$

□

11.9.4 Popoviciu's inequality for variance

Lemma 11.9.2 (Popoviciu's inequality for variance). *link Consider a random variable X with support on a finite interval $[m, M]$. Then*

its variance is bounded via

$$Var[X] = \frac{(M - m)^2}{4}.$$

- *the bound is tight and can be achieve by a discrete distribution of*

$$p(X) = \begin{cases} \frac{1}{2}, X = m \\ \frac{1}{2}, X = M \end{cases}$$

Proof. Define a function $g(t) = E[(X - t)^2]$. The derivative of g with respect to t is given by $g'(t) = -2E[X] + 2t = 0$. And the g achieves its minimum at $t = E[X]$(note that $g''(E[X]) > 0$) with minimum value $g(E[X]) = Var[X]$. Consider the special point $t = \frac{M+m}{2}$, we have

$$Var[X] = g(E[X]) \leq g(\frac{M + m}{2}) = E[(X - \frac{M + m}{2})^2].$$

Now our goal is to find an upper bound on $E[(X - \frac{M+m}{2})^2] = \frac{1}{4}E[((X-m)+(X-M))^2]$.

Since $X - m \geq 0, X - M \leq 0$, we have

$$\begin{aligned}(X-m)^2 + 2(X-m)(X-M) + (X-M)^2 &\leq (X-m)^2 - 2(X-m)(X-M) + (X-M)^2 \\ ((X-m)+(X-M))^2 &\leq ((X-m)-(X-M))^2 = (M-m)^2 \\ \implies \frac{1}{4}E[((X-m)+(X-M))^2] &\leq \frac{1}{4}E[((X-m)-(X-M))^2] = \frac{(M-m)^2}{4}.\end{aligned}$$

We therefore have

$$Var[X] = \frac{(M-m)^2}{4}.$$

□

Example 11.9.3. Consider a discrete random variable X with support on $[-1, 1]$, then the upper bound for its variance is given by

$$\frac{1}{4}(2)^2 = 1.$$

The bound can be achieve by a discrete distribution of

$$p(X) = \begin{cases} \frac{1}{2}, X = -1 \\ \frac{1}{2}, X = 1 \end{cases}$$

11.10 Convergence of random variables

11.10.1 Different levels of equivalence among random variables

Given two random variables A and B defined on the same probability space $(\Omega, \mathcal{F}, P)$, we can have the following different levels of equivalence:

- We say A is identical to B if
$$A(\omega) = B(\omega), \forall \omega \in \Omega.$$
- We say A is almost surely identical to B if
$$P(\mathcal{N}) = 0, \mathcal{N} = \{\omega, A(\omega) \neq B(\omega)\}.$$
- We say A and B have the same distribution if
$$P(A < x) = P(B < x).$$
- We say A and B have the same moments upto K if
$$E[A^k] = E[B^k], k = 1, 2, ..., K.$$

11.10.2 Convergence almost surely

We first start with the definition of almost surely convergence.

Definition 11.10.1 (convergence almost surely). *[10, p. 308]Let $\{X_n\}$ be a sequence of random variables. Then X_n converges to X almost surely if, for arbitrary $\delta > 0$ and for all $\omega \in \Omega$, we have:*
$$P(\lim_{n\to\infty} \left|X_n(\omega) - X(\omega)\right| < \delta) = 1;$$
or
$$X_n(\omega) \to X(\omega), \text{ as } n \to \infty, \forall \omega \in \Omega.$$

Intuitively, X_n converges to X almost surely if the functions $X_n(\omega)$ converges to $X(\omega)$ for all $\omega \in \Omega$ except perhaps for $s \in N, N \subset \Omega, P(N) = 0$. The probability measure o for the non-convergent point is the key point here. Note that if we view X_n as a type of function mapping, then the almost surely convergence says that X_n and X are the same (in the limit) when maps from sample space to $\mathbb{R}^n$.

Remark 11.10.1 (convergence almost surely vs. converge pointwise). If the partition the sample space Ω into two sets D and N such that $P(D) = 1$ and $P(N) = 0$. Then $X_1, X_2, ...$ converges to X almost surely is equivalently to $X_1, X_2, ...$ converges to X **pointwise** on the set of D.

Example 11.10.1. For example, let $\Omega = [0,1]$, and $X_n(\omega) = \omega + \omega^n$ and $X(\omega) = \omega$. For every $s \in [0,1)$, X_n converges to X; the non-convergent point 1 has measure of 0.

11.10.3 Convergence in probability

Definition 11.10.2 (convergence in probability). *[5]Let $\{X_n\}$ be a sequence of random variables and let X be a random variable defined on a sample space. We say that X_n converges in probability to X if, $\forall \epsilon > 0$,*

$$\lim_{n\to\infty} P(|X_n - X| >= \epsilon) = 0,$$

or equivalently

$$\lim_{n\to\infty} P(|X_n - X| < \epsilon) = 1,$$

then we write

$$X_n \xrightarrow{P} X$$

Remark 11.10.2. Note that if the random variable X is degenerate, i.e., X has close to 1 but not 1 probability of taking a constant value a. Not 1 probability means that there will infinitely often $X_n \neq a$ as $n \to \infty$. The convergence in probability is NOT like the real sequence convergence in which when n is large enough, X_n will be arbitrarily closer to a, but in probability convergence, X_n might have small chances to **take value far from** a.

Lemma 11.10.1. *Convergence almost surely will imply convergence in probability.*

Proof. Convergence almost surely says that given $\epsilon > 0$, there exist an N such that for all $n > N$, we have $\left|X_n(\omega) - X(\omega)\right| < \epsilon, \forall \omega \in A \in \mathcal{F}, P(A) \neq 0$. Therefore, $P(|X_n - X| < \epsilon) = 1, \forall n > N$, therefore, $\lim_{n\to\infty} P(|X_n - X| < \epsilon) = 1$ □

By contrast, **convergence in probability cannot imply convergence almost surely.** For example, consider $X_n(\omega) = \omega + I_{[0,1/n]}(\omega), \omega \in [0,1]$, $P_n = 1 - 1/n$ therefore it converges in probability but not almost surely since the non-convergent region has measure greater than 0.

However, **if sequence$\{X_n\}$ converges to X in probability, then there is a subsequence converges to X almost surely.**

Convergence in probability has following algebraic properties.

Theorem 11.10.1 (Algebraic properties of convergence in probability). *[5, p. 297][12, p. 1165] If $X_n \xrightarrow{P} x$ and $Y_n \xrightarrow{P} y$, then*

- $X_n + Y_n \xrightarrow{P} x + y$
- $aX_n \xrightarrow{P} ax$ *for any constant* a.
- $X_n \xrightarrow{P} x \Rightarrow g(X_n) \Rightarrow g(x)$, *for any real valued function* g *continuous at* x
- $X_n Y_n \xrightarrow{P} xy$
- $X_n / Y_n \xrightarrow{P} x/y$,*if* $y \neq 0$.
- *If* W_n *is a matrix whose elements are random variables and if* $\operatorname{plim} W_n = \Omega$, *then*

$$\operatorname{plim} W_n^{-1} = \Omega^{-1}.$$

- *If* X_n, Y_n *are random matrices with* $\operatorname{plim} X_n = A, \operatorname{plim} Y_n = B$, *then*

$$\operatorname{plim} X_n Y_n = AB.$$

Proof. (1)

$$\begin{aligned} P(\left|X_n + Y_n - x - y\right| > \epsilon) &\leq P(|X_n - x| + \left|Y_n - y\right| > \epsilon) \\ &\leq P(|X_n - x| > \epsilon/2) + P(\left|Y_n - y\right| > \epsilon/2) \to 0 \end{aligned}$$

where we have used the fact that **probability measure is monotone relative to set containment**. For the first line, $|X_n - x| + \left|Y_n - y\right| \geq \left|X_n + Y_n - x - y\right| > \epsilon$, therefore when we randomly sample X_n, Y_n, we have a higher chance to have $|X_n - x| + \left|Y_n - y\right| > \epsilon$, therefore $P(\left|X_n + Y_n - x - y\right| > \epsilon) \leq P(|X_n - x| + \left|Y_n - y\right| > \epsilon)$. For the second line, $|X_n - x| > \epsilon/2, \left|Y_n - y\right| > \epsilon/2 \Rightarrow \left|X_n + Y_n - x - y\right| > \epsilon$

(2) $P(|aX_n - ax| > \epsilon) = P(|a||X_n - x| > \epsilon) = P(|X_n - x| > \epsilon/|a|) \to 0$

(3) For any $\epsilon > 0$, there exist a δ such that $|x_n - x| < \delta \Rightarrow \left|g(x_n) - g(x)\right| < \epsilon$, therefore

$$P(\left|g(X_n) - g(x)\right| < \epsilon)) \leq P(\left|X_n) - x\right| < \delta) \to 0$$

where we have used the fact that **probability measure is monotone relative to set containment**.

(4) $X_n Y_n = \frac{1}{2} X_n^2 + \frac{1}{2} Y_n^2 - \frac{1}{2}(X_n - Y_n)^2$, use (1)(2)(3) to prove.

(5) use (3) to prove $1/Y_n \xrightarrow{P} 1/y$. (6)(7) We can approximately view matrix inversion and matrix multiplication as a series of algebraic operations on the matrix elements. □

11.10.4 Mean square convergence

Definition 11.10.3. *Let $\{X_n\}$ be a sequence of random variables. Then X_n converges to a random variable X in mean square if:*

$$\lim_{n\to\infty} E[(X_n - X)^2] = 0.$$

Theorem 11.10.2 (mean square convergence to a constant). *Let $\{X_n\}$ be a sequence of random variables and c be a constant. We say X_n converges to c if*

- $\lim_{n\to\infty} E[X_n] = c$.
- $\lim_{n\to\infty} Var[X_n] = 0$.

Proof. Use notation $\mu_n = E[X_n]$. Note that

$$\begin{aligned}
\lim_{n\to\infty} E[(X_n - c)^2] &= \lim_{n\to\infty} E[(X_n - \mu_n + \mu_n - c)^2] \\
&= \lim_{n\to\infty} E[(X_n - \mu_n)^2] + 2\lim_{n\to\infty} E[(X_n - \mu_n)(\mu_n - c)] + \lim_{n\to\infty} E[(\mu_n - c)^2] \\
&= \lim_{n\to\infty} E[(X_n - \mu_n)^2] + 0 + 0 \\
&= 0
\end{aligned}$$

□

Theorem 11.10.3 (convergence in mean square implies convergence in probability.). *Let $\{X_n\}$ be a sequence of random variables. If X_n converges to X in mean square, then X_n converges to X in probability.*

Proof. Given $\epsilon > 0$, we have

$$P(|X_n - X| > \epsilon) = P(|X_n - X|^2 > \epsilon^2) < E[(X_n - X)^2]/\epsilon^2 \to 0.$$

□

11.10.5 Convergence in distribution

Definition 11.10.4 (Convergence in distribution). *[5, p. 300]Let $\{X_n\}$ be a sequence of random variables and let X be a random variable. Let F_{X_n} and F_X be the cumulative distribution function of X_n and X. Let $C(F_X)$ denote the set of all points where F_X is continuous. We say X_n converges in distribution to X if*

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x), \forall x \in C(F_X)$$

We denote as

$$X_n \xrightarrow{D} X$$

Convergence in mean squared sense and convergence in probability can both imply convergence in distribution.

Theorem 11.10.4 (convergence in probability or in mean squared sense implies convergence in distribution). *If X_n converges to X in probability or in mean squared sense, then X_n converges to X in distribution.*

Proof. Since convergence in mean squared sense implies convergence in probability [Theorem 11.10.3], we only show convergence in probability implies convergence in distribution.

Let x be a point of continuity of $F_X(x)$. For every $\epsilon > 0$,

$$\begin{aligned} F_{X_n}(x) &= P(X_n \leq x) \\ &= P(X_n \leq x \cap |X_n - X| < \epsilon) + P(X_n \leq x \cap |X_n - X| \geq \epsilon) \\ &\leq P(X_n < x + \epsilon) + P(|X_n - X| \geq \epsilon) \end{aligned}$$

where the inequality is established by using a containing set. Then we have

$$\limsup_{n\to\infty} F_{X_n}(x) \leq P(X_n < x + \epsilon) = F_X(x + \epsilon)$$

since the second term can be arbitrarily small. Similarly, we have

$$\liminf_{n\to\infty} F_{X_n}(x) \geq P(X < x - \epsilon) = F_X(x - \epsilon)$$

We therefore have

$$F_X(x - \epsilon) \leq \liminf_{n\to\infty} F_{X_n}(x) \leq \limsup_{n\to\infty} F_{X_n}(x) \leq F_X(x + \epsilon)$$

As $\epsilon \to 0$, we have $\liminf_{n\to\infty} F_{X_n}(x) = \limsup_{n\to\infty}$ as required by the continuity of F_X, then $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$. □

Remark 11.10.3.

- In the above proof, we cannot directly use $\lim_{n\to\infty} F_{X_n}$ because it might not exist; however $\limsup_{n\to\infty} F_{X_n}$ always exists for bounded sequence.
- Convergence in distribution is weaker than convergence almost surely, because **it says nothing on the mapping from random experiment outcomes to** $\mathbb{R}$. For example, let X be a normal random variable, let $Y = -X$, then Y and X are the same in distribution, but X and Y are totally different mappings.

In general, convergence in distribution cannot imply convergence in probability. However, if X_n converges to a constant b in distribution, then X_n converges to b in probability. To see this, consider for any $\epsilon > 0$, we have $P(|X_n - b| > \epsilon) = F_{X_n}(b + \epsilon) - F_{X_n}(b - \epsilon) \to 1 - 0 = 1$.

11.11 Law of Large Number and Central Limit theorem

11.11.1 Law of Large Numbers

The Law of Large Numbers plays an essential role in applications of probability and statistics. The basic idea is quite simple: if we repeat a random experiment independently multiple times and average the result, the averaged results will be quite closed to the expectation of the random outcome. Based on the convergence mode to the expectation, there are two main versions of the law of large numbers. They are called the weak and strong laws of the large numbers. We start with the Weak Law of Large Numbers.

Theorem 11.11.1 (Weak Law of Large Numbers). *Let $\{X_n\}$ be a sequence of **iid random variables** having a common mean $E[X_i] = E[X] = \mu$ and a finite variance $\sigma^2 < \infty$. Let $\overline{X}_n = n^{-1}\sum_{i=1}^{n} X_i$. Then for every $\epsilon > 0$,*

$$\lim_{n\to\infty} P(\left|\overline{X}_n - \mu\right| < \epsilon) = 1;$$

that is, X_n converges in probability to μ.

Proof. First, we can show that $\overline{X}_n$ has the same mean of $E[X]$.

$$\begin{aligned} E[\overline{X}_n] &= \frac{EX_1 + EX_2 + \ldots + EX_n}{n} \\ &= \frac{nE[X]}{n} \\ &= E[X] \end{aligned}$$

The variance of $\overline{X}_n$ is given by

$$\begin{aligned} \text{Var}(\overline{X}) &= \frac{Var\left(X_1 + X_2 + \ldots + X_n\right)}{n^2} \\ &= \frac{Var\left(X_1\right) + Var\left(X_2\right) + \ldots + Var\left(X_n\right)}{n^2} \\ &= \frac{nVar(X)}{n^2} \\ &= \frac{Var(X)}{n}. \end{aligned}$$

We then can use Chebyshev's inequality [Theorem 11.9.1] to write

$$\begin{aligned} P(|\overline{X}_n - \mu| \geq \epsilon) &\leq \frac{Var(\overline{X}_n)}{\epsilon^2} \\ &= \frac{Var[\overline{X}_n]}{n\epsilon^2} \end{aligned}$$

which goes to zero as $n \to \infty$. □

Remark 11.11.1 (Cauchy random variable does not hold). An example where the law of large numbers does not apply is the standard Cauchy distributionLemma 12.1.41, which does not have the expectation. And the average of n such variables has the same distribution as one such variable. The probability of the averaging deviation from μ does not tend toward zero as n goes to infinity.

The Strong Low of Large Numbers, as its name suggests, gives a stronger statement on the convergence property of the expected value. Contrasting with probability convergence in the Strong Low of Large Numbers, Strong Low of Large Numbers gives almost sure convergence.

Theorem 11.11.2 (Strong Law of Large Numbers). *[7, p. 235]Let $\{X_n\}$ be a sequence of iid random variable having common mean $EX_i = \mu$,$E[|X_i|] < \infty$ and the variance $\sigma^2 < \infty$. Let $\bar{X}_n = n^{-1}\sum_{i=1}^{n} X_i$. for arbitrary $\delta > 0$:*

$$P(\lim_{n\to\infty} \left|X_n - \mu\right| < \delta) = 1$$

that is, $\bar{X}_n$ converges almost surely to μ.

Note that:

- Compared to weak law, strong law requires one more moment condition $E[|X_i|] < \infty$
- The weak law states that for a specified large n, the average $\overline{X}_n$ will be concentrated on μ. However, it may still have nonzero possibility that $\left|\overline{X}_n - \mu\right| > \epsilon$; that is, such situation will happen an infinite number of times, although at infrequent intervals.
- The strong law shows with probability 1, we have that for any $\epsilon > 0$, there exists an $N > 0$ such that the inequality $\left|\overline{X}_n - \mu\right| < \epsilon$ holds for all large enough $n > N$, except possible at zero-measure set.

11.11.2 Central limit theorem

The central limit theorem (CLT) is one of the culminations in both probability theory and probabilistic modeling. The essential result is that the sum of a large number of random variables, under certain conditions, follows approximately normal distribution.

CLT plays a central role in many real-world applications where we need to characterize the distribution of the sum of random variables even their distributions are unknown. CLT justifies the normal distribution approximation and helps overcome many practical hurdles in parameter estimation.

For example, in financial modeling, the log returns of assets are modeled by normal random variables. The CLT is also a very useful approximation tool if we like to approximate the mean and variance of the sum of a large number of random variables.

Theorem 11.11.3 (central limit theorem). *[7, p. 236][5, p. 313] Let $X_1, X_2, ..., X_n$ be a sequence* ***iid random variables*** *that have mean μ and variance $\sigma^2 < \infty$. Then the random variable*

$$Y_n = \frac{(\sum_{i=1}^n X_i/n - \mu)}{\sigma/\sqrt{n}} = \frac{(\sum_{i=1}^n X_i - n\mu)}{\sqrt{n}\sigma} = \sqrt{n}(\bar{X}_n - \mu)/\sigma$$

converges in distribution to $N(0,1)$.

Proof. Use moment generating function (if exists) or characteristic function to prove.

Let $\phi(t) = E[\exp(it(X-\mu))] = \exp(\frac{i\sigma^2t^2}{2})$ be the characteristic function of X. Then the characteristic function for Y_n can be derived via

$$\begin{aligned}
\Phi(t,n) &= E[\exp(it\frac{(\sum_{j=1}^n X_j/n - \mu)}{\sigma/\sqrt{n}})] \\
&= \phi(\frac{t}{\sigma\sqrt{n}})^n \\
&= (1 - \frac{t^2}{2n} + O((t/\sqrt{n})^3))^n \\
&\to \exp(-\frac{t^2}{2}), n \to \infty
\end{aligned}$$

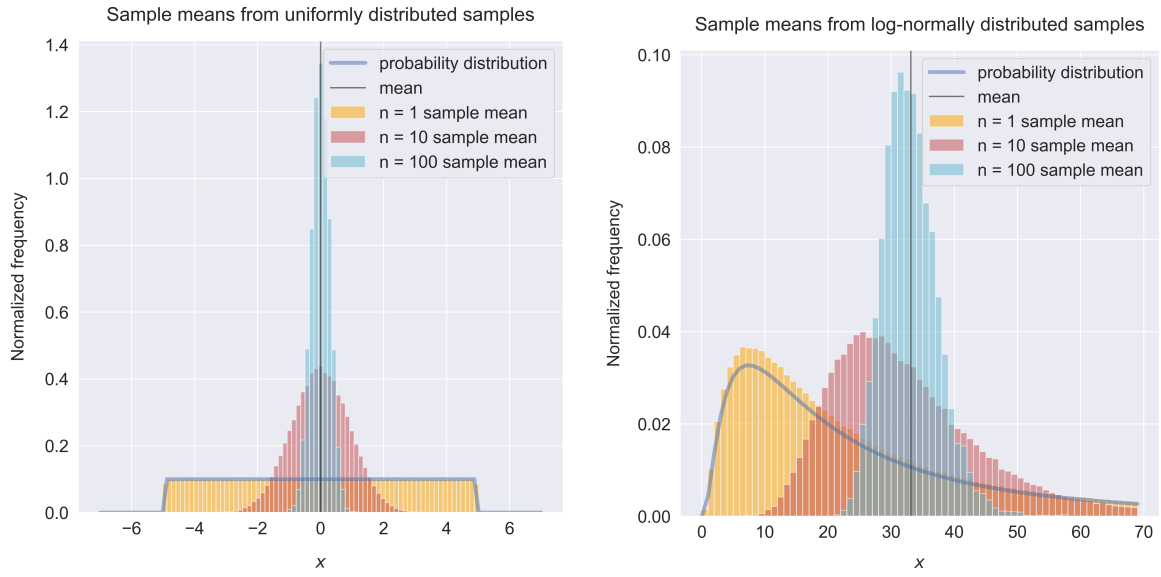
where we use the Taylor expansion of $\phi(t)$ given by

$$\phi(t) = \phi(0) + \phi'(0)t + \phi''(0)\frac{t^2}{2} + O(t^3) = 1 - \sigma^2\frac{t^2}{2} + O(t^3),$$

and the limit theorem to e [Lemma 1.5.2].

That is, as $n \to \infty$, Y_n will have its characteristic function converge to the characteristic function of the standard normal. □

Figure 11.11.1 visualizes central limit theorem for samples drawn from uniform and lognormal distributions, respectively. Sample means $\overline{X}_n$ converge to normal distribution in distribution when n is large.



(a) Samples are drawn from a uniform distribution $[-5, 5]$.

(b) Samples are drawn from a lognormal distribution $LN(1, 3)$.

Figure 11.11.1: Visualization of central limit theorem. Samples are drawn from uniform and lognormal distributions. Sample means $\overline{X}_n$ converge to normal distribution in distribution when n is large.

Remark 11.11.2 (convergence rate)**.** We can view the sample mean $\overline{X}_n$ has distribution similar to $N(\mu, \sigma/\sqrt{n})$ at large n. Therefore, the convergence rate is $O(1/\sqrt{N})$.

Remark 11.11.3 (Situations where central limit theorem breaks down)**.**

- The sample mean of the iid standard Cauchy distribution random variable will not converge in distribution to standard normal; instead, the sample mean will converge to standard Cauchy distribution [Lemma 12.1.41]. Note that standard Cauchy does not have finite mean and variance.

Finally, we give the following example to demonstrate one practical application of CLT for normal approximations.

Example 11.11.1 (application of CLT for normal approximations).

- Let $X_1, ..., X_n$ be independent iid random variable of $Exp(\lambda)$, then

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

can be approximated(when $n \to \infty$) by

$$\frac{Y - n\mu}{\sqrt{n\sigma}} \sim N(0,1),$$

where $\mu = n/\lambda$,and $\sigma = 1/\lambda^2$.

- Let $X_1, ..., X_n$ be independent iid random variable of $Poisson(\theta)$, then

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

can be approximated by

$$\frac{Y - n\theta}{\sqrt{n\theta}} \sim N(0,1),$$

or equivalently

$$Y \sim N(n\theta, \theta/n).$$

11.12 Finite sampling models

11.12.1 Counting principles

Theorem 11.12.1 (Fundamental counting principle). *Suppose that two events occur in order. If the first can occur in m ways and the second in n ways(after the first has occurred), then the two events can occur in order in $m \times n$ ways.*

Definition 11.12.1 (permutation).

- *A **permutation** of any r elements taken from a set of n elements is an arrangement of the r elements. We denote the number of such permutations by $P(n,r)$.*
- *A **permutation** is an arrangement of objects. For example, the permutations of three letters abc are the six arrangements:*

$$abc, acb, bac, bca, cab, cba.$$

Theorem 11.12.2 (number of permutations).

- *The number of permutations for n objects is*

$$P(n,n) = n!$$

.

- *The number of permutations of n objects taken from r at a time is*

$$P(n,r) = \frac{n!}{(n-r)!}$$

Proof. Choosing r elements from a set of size n, we have:

- the first element can be selected n ways.
- the second element can be selected $n-1$ ways(since now there are $n-1$ left).
- the third element can be selected $n-2$ ways.
- Continue the process, and the r^{th} element can be selected $n-r+1$ ways.

Using the fundamental counting principle [Theorem 11.12.1], we have

$$P(n,r) = n(n-1)(n-2)\cdots(n-r+1).$$

□

Lemma 11.12.1 (number of distinguishable permutations). *If a set of n objects consists of k different kinds of objects with n_i objects of the i kind such that $\sum_{i=1}^{k} = n$.* ***Objects from the same kind is not distinguishable****. Then then the number of distinguishable permutations of these objects is*

$$\frac{n!}{n_1!n_2!...n_k!}.$$

Definition 11.12.2 (combination). *A combination is a subset of elements of a set.*

Example 11.12.1. The combinations of size $r = 1, 2, 3$ taken from the set $\{a, b, c\}$ is given in the following table.

$r = 1$	$r = 2$	$r = 3$
$\{a\}$	$\{a, b\}$	$\{a, b, c\}$
$\{b\}$	$\{a, c\}$	
$\{c\}$	$\{b, c\}$	

Theorem 11.12.3 (number of combinations). *The number of combinations(or subsets) of size r which can be selected from a set of size n, denoted by $C(n, r)$ or $\binom{n}{r}$, is*

$$C(n, r) = \binom{n}{r} = \frac{n!}{r!(n-r)!}$$

Proof. Because combinations are essentially permutations where order does not matter. Then

$$C(n, r) = \frac{P(n, r)}{r!} = \frac{n!}{r!(n-r)!}.$$

□

Lemma 11.12.2 (decomposition).

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

Proof. Choosing k objects from n objects can be done by choosing k objects from $n-1$ objects or choosing $k-1$ objects from $n-1$ objects plus the rest. □

Lemma 11.12.3.

- *Given a set of n objects, the number of ways to divide them into k groups, each with n_i objects such that $\sum_{i=1}^{k} n_i = n$, is given by*

$$\frac{n!}{n_1!n_2!...n_k!}$$

- *Select n_1 objects from n objects to form a group, the number of ways is given by*

$$\frac{n!}{n_1!(n-n_1)!}.$$

Example 11.12.2. Assume 365 days a year. Among N people, the probability of exact 2 people has the same day of birthday is given as

$$365 \times \frac{1}{365} \times \frac{1}{365} \times (364 \cdot 363 \cdot \ldots (364 - (n-2) + 1)/365^{n-2}.$$

Example 11.12.3. Assume 365 days a year. Among N people, the probability of at least 2 people has the same day of birthday is given as

$$1 - \frac{365 \cdot 364 \cdots 365 - n + 1}{365^n}.$$

Example 11.12.4. 52 cards are randomly distributed to 4 players with each player getting 13 cards. What is the probability that each of them will have an ace.
Solution: The total possibilities are

$$N_0 = \frac{52!}{13!13!13!13!}.$$

The possibilities that each of them has an ace is

$$N_1 = \frac{48!}{12!12!12!12!}4!.$$

Then, we have

$$p = \frac{N_1}{N_0}.$$

Example 11.12.5. Imagine you have the following setup:

```
_A1_A2_A3_A4_
```

Each ace separated out evenly and we are interested in the pile that's before A1. For a standard deck of cards you have 52 cards - 4 aces = 48 cards left, and

$$\frac{48}{5} = 9.6,$$

cards for each pile. So basically you would have to turn all 9.6 cards + the A1 card in order to see the first ace. So the answer is

$$1 + \frac{48}{5}.$$

11.12.2 Matching problem

Example 11.12.6. A secretary randomly stuffs 5 letters into 5 envelopes. We want to find the probability of exactly k matches, with $k \in \{0, 1, ..., 5\}$.

Lemma 11.12.4 (sampling with replacement). *Define* $I_j = 1(X_j = j)$.

- $(I_1, I_2, ..., I_n)$ *is a sequence of n Bernoulli trials, with success probability* $\frac{1}{n}$.
- *The number of matches* N_n *is binomial distribution with parameter n and* $1/n$.

Lemma 11.12.5 (probability of the union of n events). *For any n events* $E_1, E_2, ..., E_n$ *that are defined on the same sample space, we have the following formula:*

$$P(E_1 \cup E_2 \cup ... \cup E_n) = \sum_{m=1}^{n} (-1)^{m+1} S_m,$$

where

$$\begin{aligned} S_1 &= \sum_{i=1}^{n} P(E_i) \\ S_2 &= \sum_{1\le j<k\le n} P(E_i \cap E_j) \\ &\cdots\cdots \\ S_m &= \sum_{1\le i_1<i_2<\ldots<i_m\le n} P(E_{i_1} \cap E_{i_2} \cap \ldots \cap E_{i_m}). \end{aligned}$$

In particular,

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2),$$

and

$$\begin{aligned} P(E_1 \cup E_2 \cup E_3) = {} & P(E_1) + P(E_2) + P(E_3) \\ & - P(E_1 \cap E_2) - P(E_1 \cap E_3) - P(E_1 \cap E_3) + P(E_1 \cap E_2 \cap E_3). \end{aligned}$$

Lemma 11.12.6 (The matching problem). *link Suppose that the n letters are numbered* $1, 2, \cdots, n$. *Let* E_i *be the event that the* i^{th} *letter is stuffed into the correct envelop.*

$P(E_1 \cup E_2 \cup \cdots \cup E_n)$ *is the probability that at least one letter is matched with the correct envelop.*

- $1 - P(E_1 \cup E_2 \cup \cdots \cup E_n)$ *is the probability that* ***all*** *letters matched incorrectly.*
- *The probability of the intersection of m events is:*

$$P(E_{i(1)} \cap E_{i(2)} \cap \cdots \cap E_{i(m)}) = \frac{(n-m)!}{n!}.$$

- $P(E_1 \cup E_2 \cup \cdots \cup E_n)$ *can be calculated using the probability of event union lemma [Lemma 11.12.5].*

Proof. (3) The calculation of the probability of intersection of m events can use the following model. There are totally $n!$ ways putting letters into envelops; there are totally $(n-m)!$ ways putting letters into envelops such that at least m specified letters are in the correct envelops. Therefore,

$$P(E_{i(1)} \cap E_{i(2)} \cap \cdots \cap E_{i(m)}) = \frac{(n-m)!}{n!}.$$

□

11.12.3 Birthday problem

Definition 11.12.3. *The sampling experiment as a distribution of n balls into m cells; X_i is the cell number of ball i. In this interpretation, our interest is in the number of empty cells and the number of occupied cells.*

Example 11.12.7. In a set of n randomly chosen people, some pair of them will have the same birthday.

Lemma 11.12.7. *Let Y_i to denote the number of balls falling into the i box, then*

$$p(Y_1 = y_1, Y_2 = y_2, \ldots, Y_m = y_m) = \frac{n!}{y_1! y_2! \ldots y_m!} \frac{1}{m^n}, \sum_{i=1}^{m} y_i = n$$

That is, the random vector $(Y_1, ..., Y_m)$ has the multinomial distribution with parameter n and $(1/m, ..., 1/m)$.

Example 11.12.8. Assume 365 days a year.

- Among N people, the probability of exact 2 people has the same day of birthday is given as

$$365 \times \frac{1}{365} \times \frac{1}{365} \times \frac{364 \cdot 363 \cdot \ldots (364 - (n-2) + 1)}{/} 365^{n-2}.$$

- The probability that at least 2 have the same birthday is

$$1 - \frac{1}{365^n} \frac{365!}{(365-n)!}.$$

Example 11.12.9. If you randomly put 18 balls into 10 boxes, what is the expected number of empty boxes? For each box, the probability of being empty is $(\frac{9}{10})^18$, then the expected number of empty boxes is $10(\frac{9}{10})^18$.

Lemma 11.12.8 (generalized birthday problem). *link Given a year with d days, the generalized birthday problem asks for the minimal number $n(d)$ such that, in a set of n*

randomly chosen people, the probability of a birthday coincidence is at least 50%. It follows that $n(d)$ is the minimal integer n such that

$$1 - (1 - \frac{1}{d})(1 - \frac{2}{d} \cdots (1 - \frac{n-1}{d})) \geq 1/2.$$

11.12.4 Coupon collection problem

Definition 11.12.4 (coupon collection problem). *Suppose that there is an urn of n different coupons. How many coupons do you expect you need to draw **with replacement** before having drawn each coupon at least once?*

Lemma 11.12.9. *Consider the coupon collection problem with m different coupons. Let Z_i denote the number of additional samples needed to go from $i-1$ distinct coupons to i distinct coupons. Let W_k denote the number of samples needed to get k distinct coupons. Then*

- *Then $Z_1, ..., Z_m$ is a sequence of independent random variables, and Z_i has the geometric distribution with parameter $p_i = \frac{m-i+1}{m}$.*
- *$W_k = \sum_{i=1}^{k} Z_i$.*
- *$E[W_k] = \sum_{i=1}^{k} \frac{m}{m-i+1}$.*

Proof. (1) When $i = 1$, Z_1 has a geometric distribution with parameter $p_1 = 1$. Similarly, Z_2 has a geometric distribution with parameter $p_2 = (m-1)/m$; Z_3 has a geometric distribution with parameter $p_3 = (m-2)/m$. Then, we can generalize to Z_i has a geometric distribution with parameter $p_i = (m-(i-1))/m$. (3) From the property of geometric distribution [Lemma 12.1.5],

$$E[W_k] = E[Z_1] + E[Z_2] + ... + E[Z_k] = \frac{1}{p_1} + \frac{1}{p_2} + ... + \frac{1}{p_k}.$$

□

Lemma 11.12.10. *Consider the coupon collection problem with m different coupons. Among m different coupons, there are $n, n \leq m$ are special coupons. Let Z_i denote the number of additional samples needed to go from $i-1$ distinct special coupons to i distinct special coupons. Let W_k denote the number of samples needed to get k distinct special coupons. Then*

- *Then $Z_1, ..., Z_m$ is a sequence of independent random variables, and Z_i has the geometric distribution with parameter $p_i = \frac{n-i+1}{m}$.*
- *$W_k = \sum_{i=1}^{k} Z_i$.*

- $E[W_k] = \sum_{i=1}^{k} \frac{m}{n-i+1}$.

Proof. (1) When $i = 1$, Z_1 has a geometric distribution with parameter $p_1 = n/m$. Similarly, Z_2 has a geometric distribution with parameter $p_2 = (n-1)/m$; Z_3 has a geometric distribution with parameter $p_3 = (n-2)/m$. Then, we can generalize to Z_i has a geometric distribution with parameter $p_i = (n-(i-1))/m$. (3) From the property of geometric distribution [Lemma 12.1.5],

$$E[W_k] = E[Z_1] + E[Z_2] + ... + E[Z_k] = \frac{1}{p_1} + \frac{1}{p_2} + ... + \frac{1}{p_k}.$$

□

11.12.5 Balls into bins model

Definition 11.12.5 (balls into bins problems). *Suppose there are m balls and n bins, balls are thrown into bins where each ball is thrown into a bin uniformly at random.*

- *Pick a bin. What is the probability for this box to be empty? What is the expected number of bins that are empty?*
- *Pick a bin. What is the probability for this box to contain exactly 1 ball? What is the expected number of bins that contain exactly 1 ball.*
- *Pick a bin. What is the probability for this box to contain exactly i balls? What is the expected number of bins that contain exactly i balls?*

Example 11.12.10. Suppose there are N types of coupons in a box. If a child draws with replacement m times from the box, what is the expected number of distinct coupon types?

View each coupon as a box. And this problem is equivalent to throw m balls into N boxes and ask the expected number of non-empty boxes.

- For each box, the probability of being empty is $(\frac{N-1}{N})^m$; therefore, the probability of being non-empty is $1 - (\frac{N-1}{N})^m$.
- The expected number of empty boxes is $N(\frac{N-1}{N})^m$, and nonempty boxes is $N - N(\frac{N-1}{N})^m$.

Definition 11.12.6 (balls-into-bins distribution problems).

- *(distribution of distinguishable balls into indistinguishable bins without restriction) Suppose we want to put m distinguishable balls into n labeled bins. What is the number of ways that the balls are in different bins?*
- *(distribution of indistinguishable balls into indistinguishable bins without restriction) Suppose we want to put m indistinguishable balls into n labeled bins. What is the number of ways that the balls are in different bins?*
- *(distribution of distinguishable balls into distinguishable bins without restriction) Suppose we want to put m indistinguishable balls into n labeled bins. What is the number of ways that the balls are in different bins?*
- *(distribution of indistinguishable balls into distinguishable bins without restriction) Suppose we want to put m indistinguishable balls into n labeled bins. What is the number of ways that the balls are in different bins?*
- *(distribution without restriction I) Suppose we want to put m labeled balls into n labeled bins. What is the number of ways that the balls are in different bins such that each bin has at least has one ball?*
- *(distribution without restriction II) Suppose we want to put m labeled balls into n labeled bins. What is the number of ways that the balls are in different bins such that each bin has at least k balls?*

Lemma 11.12.11. *link*

The number of ways of putting m distinguishable balls in n distinguishable bins is m^n.

- *The number of ways of putting m distinguishable balls in n indistinguishable bins is $m^n/n!$.*
- *The number of ways of putting m indistinguishable balls in n distinguishable bins is*

$$\binom{m+n-1}{n-1}.$$

- *The number of ways of putting m indistinguishable balls in n indistinguishable bins is*

$$\binom{m+n-1}{n-1}\frac{1}{n!}.$$

- *The number of ways of putting m indistinguishable balls in n distinguishable bins and ensure each bin has at least one ball is*

$$\binom{(m-n)+n-1}{n-1}.$$

Proof. (1) The number of ways of putting m balls in n bins is m^n since each ball has n bins to go. (2) Use (1) and divide it by double counting. (3, 4) Transform to selecting $n-1$ separations from $m+n-1$ possibilities. (5) We need to first put one ball into each bin. □

Remark 11.12.1 (equivalence of distinct root problem)**.**

- The number of ways to distribute m indistinguishable balls into n distinguishable bins is equivalent to the number of solutions to the equation:

$$x_1 + x_2 + x_3 + \ldots + x_n = m, x_i \geq 0.$$

- The number of ways to distribute m indistinguishable balls into n distinguishable bins and ensure each bin to have at least one ball is equivalent to the number of non-negative solutions to the equation:

$$x_1 + x_2 + x_3 + \ldots + x_n = m, x_i \geq 1.$$

Example 11.12.11. If there are 200 students in the library, how many ways are there for them to be split among the floors of the library if there are 6 floors? The answer is 6^200.

11.13 Order statistics

Definition 11.13.1. *The order statistics of a random sample $X_1, ..., X_n$ are the sample values placed in ascending order. And they are denoted by $X_{(1)}, X_{(n)}, ..., X_{(n)}$, where $X_{(1)} \leq X_{(2)} ... \leq X_{(n)}$*

Theorem 11.13.1 (Discrete order statistics). *[7] Let $X_1, ..., X_n$ be a random sample from a discrete distribution with pmf $f_X(x_i) = p_i$, where $x_1 < x_2 < ...$ are possible values of X in ascending order. Define*

$$\begin{aligned} P_0 &= 0 \\ P_1 &= p_1 \\ &... \\ P_i &= \sum_{k=0}^{i} p_i \end{aligned}$$

Let $X_{(1)}, X_{(n)}, ..., X_{(n)}$ denote the order statistics from the sample. Then

$$P(X_{(j)} \leq x_i) = \sum_{k=j}^{n} \binom{n}{k} P_i^k (1 - P_i)^{n-k} \tag{6}$$

and

$$P(X_{(j)} = x_i) = \sum_{k=j}^{n} \binom{n}{n-k} [P_i^k (1 - P_i)^{n-k} - P_{i-1}^k (1 - P_{i-1})^{n-k} \tag{7}$$

Proof. We can treat P_i as discrete version of cdf, and it means the probability of one X satisfies the inequality. The order statistics connected to binomial distribution as:

- If the minimum of Xs are less then x, then there are 1,2,...,n out of n are less than x.
- If the second minimum of Xs are less then x, then there are 2,3,...,n out of n are less than x.

□

Theorem 11.13.2 (Continuous order statistics). *[7] Let $X_1, ..., X_n$ be a random sample from a continuous distribution with pmf f_X and cdf $F_X(x)$. Let $X_{(1)}, X_{(n)}, ..., X_{(n)}$ denote the order statistics from the sample. Then*

$$f_{X_{(j)}}(x) = \frac{n!}{(j-1)!(n-j)!} f_X(x)[F_X(x)]^{j-1}[1-F_X(x)]^{n-j}$$

Proof. We can use

$$F_{X_{(j)}}(x) = \sum_{k=j}^{n} \binom{n}{k} [F_X(x)]^k (1-F_X(x))^{n-k}$$

and take derivative.
Another proof: When jth order statistic at x, that means we from n variables, we first select 1 variable to be at x, then from rest of $n-1$ variables, we select $j-1$ to be smaller than x then the rest greater than x. From combinatorics, we know

$$f_j(x) = nf(x)\binom{n-1}{j-1}(F(x))^{j-1}(1-F(x))^{n-j}$$

□

Lemma 11.13.1 (Two order statistics). *Let $X_1, ..., X_n$ be a random sample from a continuous distribution with pmf f and cdf $F(x)$. Let $X_{(1)}, X_{(n)}, ..., X_{(n)}$ denote the order statistics from the sample. Then the joint density for $X_{(r)}$ and $X_{(s)}$ is:*

$$f_{r,s}(u,v) = \frac{n!}{(r-1)!(n-s)!(s-r-1)!} f(u)f(v)(F(u))^{r-1}(1-F(v))^{n-s}(F(v)-F(u))^{s-r-1}$$

Proof. Use the argument similar to above: just divide the variables into five groups. □

Corollary 11.13.2.1. *Let $X_1, ..., X_n$ be a random sample from a continuous distribution with pmf f and cdf $F(x)$. Let $X_{(1)}, X_{(n)}, ..., X_{(n)}$ denote the order statistics from the sample. Then we have*

- $$f_{1,n}(u,v) = n(n-1)(F(v)-F(u))^{n-2}f(u)f(v)$$
- *(density of range) Let $W = X_{max} - X_{min}$, then*

$$f_W(w) = \int_u f_{1,n}(u, u+w)du$$

Lemma 11.13.2 (joint density of all the order statistics). *Let $X_1, ..., X_n$ be a random sample from a continuous distribution with pmf f and cdf $F(x)$. Let $X_{(1)}, X_{(n)}, ..., X_{(n)}$ denote the order statistics from the sample. Then the conditional joint density function of $X_{(1)}, X_{(n)}, ..., X_{(n)}$ is given by*

$$f_{1,2,...,n}(y_1, ..., y_n) = n! f(y_1) f(y_2) ... f(y_n) I_{y_1 < y_2 ... < y_n}$$

Proof. The sample space of $(X_1, ..., X_n)$ can be partition into $n!$ **equal-sized** subspaces such that $X_1 < X_2 < ... < X_n, ...$ In each this subspace, there exist a map from $(X_1, ..., X_n)$ to $(X_{(1)}, X_{(n)}, ..., X_{(n)})$ with the Jacobian being 1(since it is a permutation matrix). The density for $(X_{(1)}, X_{(n)}, ..., X_{(n)})$ is $f(y_1) f(y_2) ... f(y_n) I_{y_1 < y_2 ... < y_n}$. Use the law of total probability [Theorem 11.2.1]. □

Lemma 11.13.3 (distribution of max and min). *Let X be a random variable with cdf $F_X(x)$. Let $Y_n = \min(X_1, ..., X_n)$ and $Z_n = \max(X_1, ..., X_n)$, where $X_1, ..., X_n$ are n iid random sample of X. Then*

$$f_{Y_n}(x) = n f_X(x)(1 - F_X(x))^{n-1}$$

and

$$f_{Z_n}(x) = n f_X(x)(F_X(x))^{n-1}.$$

Proof.

$$P(Y_n \geq x) = (P(X \geq x))^n \implies 1 - F_{Y_n}(x) = (1 - F_X(x))^n \implies f_{Y_n}(x) = n f_X(x)(1 - F_X(x))^{n-1}$$

and

$$P(Z_n \leq x) = (P(X \leq x))^n \implies F_{Z_n}(x) = (F_X(x))^n \implies f_{Z_n}(x) = n f_X(x)(F_X(x))^{n-1}.$$

□

Corollary 11.13.2.2 (order statistics of uniform random variables). *Let X be a uniform random variable at [o,1]. Let $Y_n = \min(X_1, ..., X_n)$ and $Z_n = \max(X_1, ..., X_n)$, where $X_1, ..., X_n$ are n iid random sample of X. Then*

- $$f_{Y_n}(x) = n f_X(x)(1 - F_X(x))^{n-1} = n(1 - x)^{n-1}$$
- $$f_{Z_n}(x) = n f_X(x)(F_X(x))^{n-1} = n x^{n-1}$$

- $$f_j = \frac{n!}{(j-1)!(n-j)!}[x]^{j-1}[1-x]^{n-j} = Beta(j, n-j+1)$$

Proof. Note that we use the fact that for $U(0,1)$ distribution, $F_X(x) = x, f_X(x) = 1$. □

11.14 Information theory

11.14.1 Concept of entropy

Definition 11.14.1 (entropy of a random variable).

- *Let X be a discrete random variable taking values $x_k, k = 1, 2...$ with probability mass function*
$$P(X = x_k) = p_k, k = 1, 2, ...$$
Then the entropy of X is defined by
$$H(X) = -\sum_{k\geq 1} p_k \ln p_k.$$
- *If X is a continuous random variable with pdf $f(x)$, then entropy of X is defined by*
$$H(X) = -\int_{-\infty}^{\infty} f(x) \ln f(x) dx.$$

Remark 11.14.1 (entropy, information and probability distribution).

- Entropy is a measure of the uncertainty of a random variable: the larger the value, the uncertainty the random variable is.
- When the random variable is deterministic, the entropy is at the minimum.

Example 11.14.1.

- The entropy of the Gaussian density on $\mathbb{R}$ with mean μ and variance σ^2 is
$$\begin{aligned} H &= -\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}\sigma} \exp(-1/2((x-\mu)^2/\sigma^2))(-\ln(\sqrt{2\pi}\sigma) - 1/2((x-\mu)^2/\sigma^2))dx \\ &= \frac{1}{2} + \ln(\sqrt{2\pi}\sigma). \end{aligned}$$
Note that the mean μ does not enter the entropy; therefore the entropy for Gaussian distribution is translational invariant.
- The entropy of the exponential distribution with mean λ and pdf
$$f(x) = \frac{1}{\lambda} \exp(-x/\lambda)$$

is

$$H = -\int_0^\infty \frac{1}{\lambda}\exp(-x/\lambda)(-\ln\lambda - x/\lambda)dx = \ln\lambda + 1.$$

Lemma 11.14.1 (basic properties of entropy).

- $H(X) \geq 0$.
- $H(X) = 0$ *if and only if there exists a* x_0 *such that* $P(X = x_0) = 1$.
- *If* X *can take on finite number* n *values, then* $H(X) \leq \log(n)$. $H(X) = \log(n)$ *if and only if* X *is uniformly distributed.*
- *Let* $X_1, X_2, ..., X_n$ *be discrete valued random variables on a common probability space. Then*
$$H(X_1, X_2, ..., X_n) = H(X_1) + \sum_{i=2}^{n} H(X_n|X_1, ..., X_{n-1}).$$
- $H(X) + H(Y) \geq H(X, Y)$, *with equality if and only if* X *and* Y *are independent.*

Proof. (1) note that every term $\log(p)$ is non-positive, therefore $H(X) \geq 0$. (2) direct verification. (3) direct verification. (4) It can be showed that $H(X, Y) = H(X|Y) + H(Y)$. (5) $H(X, Y) = H(X|Y) + H(Y) \leq H(X) + H(Y)$(using chain rule and conditioning entropy). □

11.14.2 Entropy maximizing distributions

Theorem 11.14.1 (continuous distribution with maximum entropy). *Suppose* S *is a closed subset of* $\mathbb{R}$. *Let* X *be a random variable with support* S *and pdf* $f(x)$.

Then, the probability density function $f(x)$ *maximizing the entropy*

$$H(X) = -\int_{-\infty}^{\infty} f(x)\ln f(x)dx,$$

and satisfying the following n *constraints*

$$E[g_j(X)] = a_j, \forall j = 1, 2, ..., n.$$

and sum-to-unit constraint

$$\int_S f(x)dx = 1,$$

has the form

$$f(x) = c\exp(\sum_{j=1}^{n} \lambda_j g_j(x)), \forall x \in S,$$

where the constant c and the n multipliers λ_i are determined by the above $n+1$ constraints.

Proof. Note that our constraints can be written as

$$\int_{-\infty}^{\infty} g_j(x) f(x) dx = a_j, j = 1, 2, ..., n$$

$$\int_{-\infty}^{\infty} f(x) dx = 1, j = 1, 2, ..., n.$$

The Lagrange of our minimizing problem is given by

$$J[p(x)] = \int_{-\infty}^{\infty} f(x) \ln f(x) dx - \lambda_0 (\int_{-\infty}^{\infty} f(x) dx - 1) - \sum_{j=1}^{n} \lambda_j (\int_{-\infty}^{\infty} g_j(x) f(x) dx - a_j).$$

where $\lambda_i, i = 0, 1, 2, ..., n$ are Lagrange multipliers.

The first order optimality condition gives

$$\frac{\delta J}{\delta f(x)} = \ln f(x) + 1 - \lambda_0 - \lambda_j g_j(x),$$

or equivalently

$$f(x) = \exp(-1 + \lambda_0) \exp(\sum_{j=1}^{n} \lambda_j g_j(x)) = c\exp(\sum_{j=1}^{n} \lambda_j g_j(x)).$$

Note that the second order conditions gives $\frac{\delta^2 J}{\delta f(x)^2} = 1/f(x) > 0$, which ensures we have unique global minimum solution. □

Corollary 11.14.1.1.

- *The uniform distribution on the interval $[a, b]$ is the maximum entropy distribution among all continuous distribution supported on $[a, b]$.*
- *The exponential distribution, for which the density function with parameter λ is*

$$f(x|\lambda) = \begin{cases} \lambda \exp(-\lambda x), x \geq 0 \\ 0, x < 0 \end{cases}$$

is the maximum entropy distribution among all continuous distributions supported in $[0, \infty]$ that have a specified mean of $1/\lambda$.

- *The normal distribution with parameter μ and σ, for which the density function is*

$$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(x-\mu)^2}{2\sigma^2}),$$

has the maximum entropy among all distributions supported on $\mathbb{R}$ with a specified mean μ and variance σ^2.

Proof. (1) from Theorem 11.14.1, we know that the $f(x)$ should have the following form

$$f(x) = c.$$

and c is determined by

$$\int_a^b f(x)dx = c(b-a) = 1 \implies c = \frac{1}{b-a}.$$

Therefore,

$$f(x) = \frac{1}{b-a}, x \in [a,b].$$

(2)Similarly, we know that the $f(x)$ should have the following form

$$f(x) = c\exp(\mu x),$$

where μ is the Lagrange multiplier and c is determined by

$$\int_0^\infty f(x)dx = \frac{c}{\mu} = 1 \implies c = \mu.$$

and then

$$\int_0^\infty xf(x)dx = -\frac{1}{\mu} = 1/\lambda \implies \mu = -\lambda.$$

(3)Similarly, we know that the $f(x)$ should have the following form

$$f(x) = c\exp(\lambda_1 x + \lambda_2(x-\mu)^2).$$

Then we can determine c, λ_1, λ_2 using constraints. □

Theorem 11.14.2 (discrete distribution with maximum entropy). *Suppose $S = \{x_1, x_2, ...\}$ is a (finite or infinite) discrete subset of of $\mathbb{R}$. Let X be a random variable with support S and probability mass function given by $P(X = x_k)$.*

Then, the probability mass function $P(X)$ maximizing the entropy

$$H(X) = -\sum_{k\geq 1} P(X = x_k)\ln P(X = x_k),$$

and satisfying the following n constraints

$$E[g_j(X)] = a_j, \forall j = 1, 2, ..., n.$$

and sum-to-unit constraint

$$\sum_{k \geq 1} P(X = x_k) = 1,$$

has the form

$$P(X = x_k) = c \exp(\sum_{j=1}^{n} \lambda_j g_j(x_k)), \forall x_k \in S,$$

where the constant c and the n multipliers λ_i are determined by the above $n + 1$ constraints.

Proof. Note that our constraints can be written as

$$\int_{-\infty}^{\infty} g_j(x) f(x) dx = a_j, j = 1, 2, ..., n$$

$$\int_{-\infty}^{\infty} f(x) dx = 1, j = 1, 2, ..., n.$$

Let $p_k = P(X = x_k)$. The Lagrange of our minimizing problem is given by

$$L = \sum_{i=1}^{n} p_k \ln p_k - \lambda_0 (\sum_{i \geq 1} p_k - 1) - \sum_{j=1}^{n} \lambda_j (\sum_{i \geq 1} g_j(x_i) p_i - a_j).$$

where $\lambda_i, i = 0, 1, 2, ..., n$ are Lagrange multipliers.

The first order optimality condition for p_i gives

$$\frac{\partial L}{\partial p_i} = \ln p_i + 1 - \lambda_0 - \lambda_j g_j(x_i), i \geq 1$$

or equivalently

$$P(X = x_i) = p_i = \exp(-1 + \lambda_0) \exp(\sum_{j=1}^{n} \lambda_j g_j(x_i)) = c \exp(\sum_{j=1}^{n} \lambda_j g_j(x_i)).$$

Note that the second order conditions gives $\frac{\partial^2 L}{\partial p_i} = 1/p_i > 0$, which ensures we have unique global minimum solution. □

Corollary 11.14.2.1. *For a probabilistic mass function p on a finite set $\{x_1, x_2, ..., x_n\}$, the entropy H is bounded by*

$$H \leq \ln n$$

with equality holds if and only if p is uniform, i.e., $p(x_i) = 1/n, \forall i$.

Proof. From Theorem 11.14.2, we know that the $p(x)$ should have the following form

$$p(x_i) = c.$$

and c is determined by

$$\sum_{i=1}^{n} c = 1 \implies c = \frac{1}{n}.$$

Therefore,

$$p(x_i) = 1/n, \forall i.$$

□

11.14.3 KL divergence

Definition 11.14.2 (Kullback-Leibler divergence, KL divergence). *Given two discrete probability distribution P and Q defined on the same set $\mathcal{X}$, the KL divergence from Q to P is defined as*

$$D_{KL}(P||Q) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}$$

Lemma 11.14.2 (non-negativeness of KL divergence). *Given two discrete probability distribution P and Q defined on the same set $\mathcal{X}$,*

$$D_{KL}(P||Q) \geq 0.$$

And the equality holds if $P = Q$.

Proof.

$$D_{KL}(P||Q) = -\sum_{x \in \mathcal{X}} P(x) \log \frac{Q(x)}{P(x)} \geq -\log(\sum_{x \in \mathcal{X}} \frac{Q(x)}{P(x)} P(x)) = 0$$

where the fact that $-\log(x)$ is a convex function and Jensen's inequality has been used [Lemma 11.9.1]. □

11.14.4 Conditional entropy and mutual information

Definition 11.14.3 (conditional entropy).

- ***Specific conditional entropy*** $H(X|Y=v)$ *of* X *given* $Y=v$*:*

$$H(X|Y=v) = -\sum_{i=1}^{n} P(X=i|Y=v)\log P(X=i|Y=v).$$

- ***Conditional entropy*** $H(X|Y)$ *of* X *given* Y*:*

$$H(X|Y) = \sum_{v\in Val(Y)} P(Y=v)H(X|Y=v).$$

Definition 11.14.4 (mutual information). *[13] Consider two discrete random variables* X *and* Y *taking values in* $\mathcal{X}$ *and* $\mathcal{Y}$*. The* ***mutual information, or information gain)*** *of* X *and* Y *is given as:* $I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$

Lemma 11.14.3.

$$I(X,Y) \geq 0$$

where $I(X,Y) = 0$ *if* X *and* Y *are independent.*

Proof. (1)

$$\begin{aligned} I(X,Y) &= H(X) - H(X|Y) \\ &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log(p(x)) + \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x|y)p(y)\log(p(x|y)) \\ &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log(p(x)) + \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log(\frac{p(x,y)}{p(y)}) \\ &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log(\frac{p(x,y)}{p(x)p(y)}) = DL(p(x,y)||p(x)p(y)) \geq 0. \end{aligned}$$

(2) When X and Y are independent, we have

$$H(X|Y) = \sum_{v\in\mathcal{Y}} (Y=v)H(X|Y=v) = \sum_{v\in\mathcal{Y}} P(Y=v)H(X) = H(x).$$

□

Corollary 11.14.2.2 (conditioning reduce entropy). *Given discrete random variables X and Y, we have*

$$H(X|Y) \leq H(X),$$

which is also known as conditioning reduces entropy (i.e., conditioning provides information); and this equality holds if and only if X and Y are independent.

Chain rule: $H(X,Y) = H(X) + H(Y|X)$ can be proved using $P(X,Y) = P(X|Y)P(Y)$.

11.14.5 Cross-entropy

Definition 11.14.5 (cross-entropy of two probability distributions). *Consider a probability distribution on N value with probability $y_i, i = 1, 2, ..., N$. Consider another distribution on the same support and probabilities $y'_i, i = 1, 2, ..., N$. Then the cross-entropy of the two distributions is defined by*

$$H(y, y') = \sum_{i=1}^{N} y_i \log \frac{1}{y'_i} = -\sum_{i=1}^{N} y_i \log y'_i.$$

Lemma 11.14.4 (properties of cross entropy). *Consider two discrete distributions, characterized by probability mass vectors y and y', on the same N values.*

- *The KL divergence on the two distributions is the difference between cross entropy and entropy; that is*

$$KL(y||y') = \sum_{i=1}^{N} y_i \log \frac{y_i}{y'_i} = \underbrace{\sum_{i=1}^{N} y_i \log \frac{1}{y'_i}}_{cross\ entropy} - \underbrace{\sum_{i=1}^{N} y_i \log \frac{1}{y_i}}_{entropy}.$$

- *Cross entropy is no smaller than entropy*

$$H(y, y') \geq H(y).$$

Proof. (1) Straight forward. (2) Use the fact that

$$KL(y||y') = H(y, y') - H(y) \geq 0.$$

□

Remark 11.14.2 (cross entropy, maximum likelihood, and classificiation accuracy). Consider a K-class classification problem with N training examples. The target of each example is represented by a K-dimensional one-hot vector. The classification output generated by the classifier can be represented by a discrete distribution vector.

For example, let $y^{(1)} = (1, 0, 0,)$ be the target vector of example 1 and $\hat{y}^{(1)} = (0.4, 0.1, 0.5)$ be a prediction output based on input of example 1.

Note that the likelihood for example i is given by

$$L(y^{(i)}; \hat{y}^{(i)}) = \prod_{k=1}^{K} [\hat{y}_j^{(i)}]^{y^{(i)}},$$

whose the logarithm form is

$$\log L(y^{(i)}; \hat{y}^{(i)}) = \sum_{k=1}^{K} y_k^{(i)} \log \hat{y}_k^{(i)} = -H(y^{(i)}, \hat{y}^{(i)}).$$

For overall N examples, the overall negative log likelihood is

$$-\log L = -\sum_{n=1}^{N} \log L(y^{(n)}; \hat{y}^{(n)}) = \sum_{n=1}^{N} H(y^{(i)}, \hat{y}^{(i)}).$$

Therefore, **minimizing the negative log likelihood is equivalent to minimizing the cross-entropy.**

11.15 Notes on bibliography

For excellent treatment on the whole topic, see[14][15]. For clear treatment on conditional expectation, see [16],[9].

For clear treatment on σ field and measure, see[1][17].

For problems in probability, see [18][19].

For treatment on measure and integral, see [20].

An excellent online resource is http://www.math.uah.edu/stat/, including random variable vector space theory(http://www.math.uah.edu/stat/expect/Spaces.html), finite sampling model(http://www.math.uah.edu/stat/urn/index.html), Brownian motion (http://www.math.uah.edu/stat/brown/Standard.html).

BIBLIOGRAPHY

1. Dineen, S. *Probability theory in finance: a mathematical guide to the Black-Scholes formula* (American Mathematical Soc., 2013).
2. Rosenthal, J. S. *A first look at rigorous probability theory* (World Scientific, 2006).
3. Wikipedia. *Borel set — Wikipedia, The Free Encyclopedia* [Online; accessed 18-May-2016]. 2016.
4. Shreve, S. E. *Stochastic calculus for finance II: Continuous-time models* (Springer Science & Business Media, 2004).
5. Hogg, R. V., McKean, J. & Craig, A. T. Introduction to Mathematical Statistics, 7 ed (2012).
6. Fries, C. *Mathematical finance: theory, modeling, implementation* (John Wiley & Sons, 2007).
7. Casella, G. & Berger, R. L. *Statistical inference* (Duxbury Pacific Grove, CA, 2002).
8. Williams, D. *Probability with martingales* (Cambridge university press, 1991).
9. Mikosch, T. *Elementary stochastic calculus with finance in view* (World scientific, 1998).
10. Grimmett, G. & Stirzaker, D. *Probability and Random Processes* ISBN: 9780198572220 (OUP Oxford, 2001).
11. Tripathi, G. A matrix extension of the Cauchy-Schwarz inequality. *Economics Letters* **63,** 1–3 (1999).
12. Greene, W. *Econometric Analysis* ISBN: 9780134461366 (Pearson, 2017).
13. Cover, T. M. & Thomas, J. A. *Elements of information theory* (John Wiley & Sons, 2012).
14. Shiryaev, A. N. Probability: Volume 1 (Graduate Texts in Mathematics) (1996).
15. Feller, W. *An introduction to probability theory and its applications* (John Wiley & Sons, 2008).
16. Brzezniak, Z. & Zastawniak, T. *Basic stochastic processes: a course through exercises* (Springer Science & Business Media, 1999).
17. Koralov, L. & Sinai, Y. G. *Theory of probability and random processes* (Springer Science & Business Media, 2007).
18. Capinski, M. & Zastawniak, T. J. *Probability through problems* (Springer Science & Business Media, 2013).

19. Grimmett, G. & Stirzaker, D. *One thousand exercises in probability* (Oxford University Press, 2001).

20. Capinski, M. & Kopp, P. E. *Measure, integral and probability* (Springer Science & Business Media, 2013).