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.*

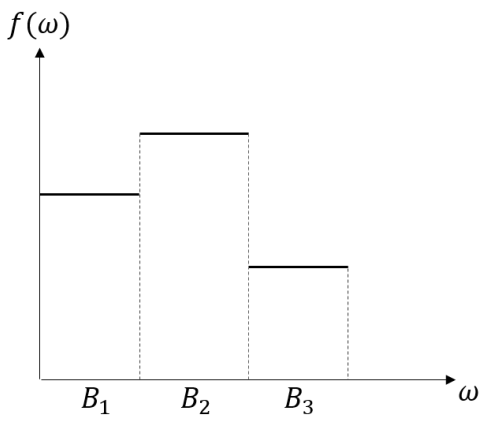
Example 11.1.3 (probability measure). A **probability measure** is a measure satisfying above three properties and has one additional requirement of total measure one $\mu(X) = 1$.

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 $f^{-1}(B) \in \mathcal{F}, B \in \mathcal{B}(\mathbb{R})$.*

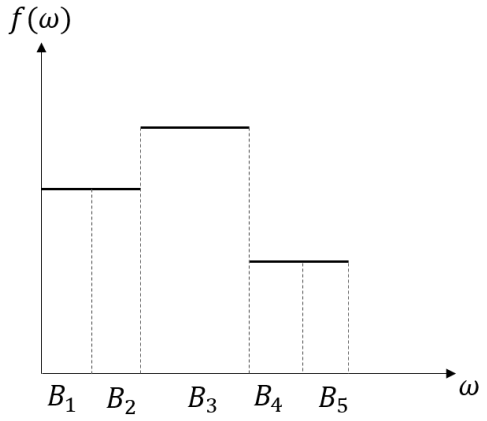
Example 11.1.4 (measurable function with coarse sigma field). Let $\mathcal{F}$ generated by a finite partitions $B_1, B_2, ..., B_m$ of Ω; let function $f : \Omega \to \mathbb{R}$ be **$\mathcal{F}$-measurable**. Then f take constant value on each element of $B_i, 1 \leq i \leq m$ [Figure 11.1.1].

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.

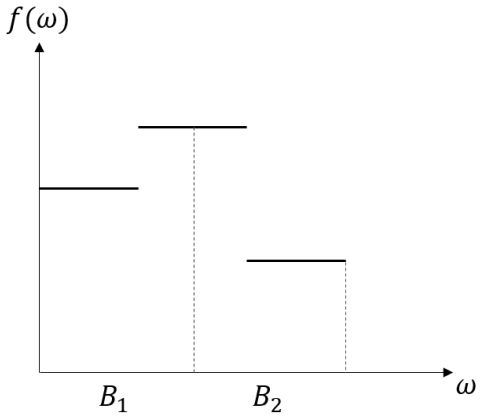
Therefore measurability usually limits the 'variation' of a function defined on a set.



(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

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

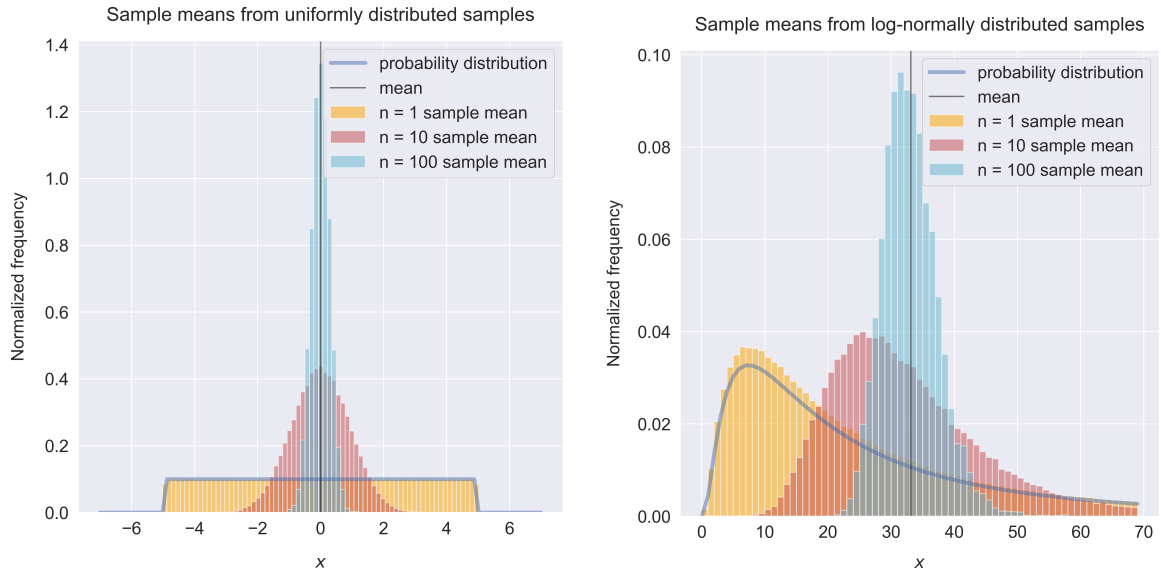
where we use the Taylor expansion of $\phi(t)$ given by

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

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).

12

STATISTICAL DISTRIBUTIONS

12.1 Common distributions and properties

12.1.1 Overview

In this chapter we survey statistical distributions commonly used in real-world applications and in subsequent chapters of this book. For each distribution, we study their basic properties like its mean and variance, and many other useful properties allowing us to construct more complex distributions. Along the lines, we will also discuss the connections between different statistical distributions.

We do not pretend to give a comprehensive overview on all aspects of statistical distributions. For an extensive discussion on statistical distributions, see [1][2].

12.1.2 Bernoulli distribution

Definition 12.1.1 (Bernoulli distribution). *A random variable Y with sample space $\{0,1\}$ is said to have Bernoulli distribution $Ber(\theta)$ with parameter θ if it has a pmf given as*

$$p(y) = \theta^y (1-\theta)^{1-y}, y \in \{0,1\}.$$

Example 12.1.1. 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.

Lemma 12.1.1 (basic properties). *Let X be a random variable with distribution $Ber(p)$. Then*

- $M_X(t) = (1 - p + pe^t)$.
- $E[X] = p, E[X^2] = p, Var[X^2] = p - p^2 = p(1-p)$.

Proof. Straight forward. □

12.1.3 Poisson distribution

Definition 12.1.2 (Poisson distribution). *A discrete random variable X is said to have a Poisson distribution $Poisson(\lambda)$ with parameter λ if it has pmf given as*

$$p(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

with $x \in \{0, 1, 2, ...\}$.

Lemma 12.1.2 (basic property of Poisson distribution). *[3, p. 154] Let X be a random variable with distribution $Poisson(\lambda)$. Then*

- $M(t) = \exp(\lambda(e^t - 1))$.
- $E[X] = \lambda, Var[X] = \lambda$.

Proof. (1)

$$\begin{aligned} M_X(t) &= E[e^{tX}] \\ &= \sum_{n=0}^{\infty} e^{tn} \frac{\lambda^n}{n!} e^{-\lambda} \\ &= e^{-\lambda} \sum_{n=0}^{\infty} \frac{(\lambda e^t)^n}{n!} \\ &= e^{-\lambda} e^{\lambda e^t}. \end{aligned}$$

(2) $E[X] = M'_X(0) = \lambda, E[X^2] = M''_X(0) = \lambda^2 - \lambda$. □

Lemma 12.1.3 (sum of Poisson distribution). *Assume $X_1, ..., X_n$ to be independent random variables, and $X_i \sim Poisson(\theta_i), i = 1, ..., n$. Then*

$$Y = \sum_{i=1}^{n} X_i \sim Poisson(\sum_{i=1}^{n} \theta_i).$$

Proof. Note that

$$M_Y(t) = \prod_{i=1}^{n} M_{X_i}(t) = \exp \sum_{i=1}^{n} \theta_i (e^t - 1).$$

□

Lemma 12.1.4 (Normal approximate sum of Poisson). *Let $X_1, ..., X_n$ be independent iid random variable of Poisson(θ), 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

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

Proof. Directly from Central Limit Theorem [Theorem 11.11.3]. □

12.1.4 Geometric distribution

Definition 12.1.3 (geometric distribution). *A discrete random variable X is said to have a geometric distribution Geo(θ) with parameter θ if it has pmf given as*

$$p(X = k) = (1 - \theta)^{k-1}\theta$$

with $k \in \{1, 2, ...\}$.

Example 12.1.2 (number of trials needed to succeed in Bernoulli trias). The geometric distribution The probability distribution of the number X of Bernoulli trials needed to get one success, supported on the set $\{1, 2, 3, ...\}$

Lemma 12.1.5 (basic statistics of geometric distribution). *The expected value of a geometrically distributed random variable X with parameter p is $1/p$ and the variance is $(1 - p)/p^2$.*

Proof. (1)

$$E[X] = \sum_{k=1}^{\infty} k(1 - p)^{k-1}p$$

$$(1 - p)E[X] = \sum_{k=1}^{\infty} k(1 - p)^{k}p$$

subtract and get $pE[X] = 1$.
(2)

$$Var[X] = \sum_{k=1}^{\infty}(k - 1/p)^2(1-p)^{k-1}p$$

can be proved similarly. □

Example 12.1.3 (coupon collection problem, subsection 11.12.4). Consider the **coupon collection problem** where there is an urn of m different coupons. How many coupons do you expect you need to draw **with replacement** before having drawn each coupon at least once?

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 $Z_j, j = 1, ..., m$ is a sequence of independent random variables has the geometric distribution with parameter $p_i = \frac{m-i+1}{m}$.

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 $p_i = (m-(i-1))/m$.

Further, we have

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

Example 12.1.4 (number of visits in a Markov chain, Lemma 20.2.1). Consider a state i in a Markov chain. Let f_{ii} denote the probability that a trajectory starting from state i will *ever* revisit i.

Then the probability of of n visit is $f_{ii}^{n-1}(1-f_{ii})$, which is the product of the probability visiting state i $n-1$ times and then never visit again.

The expected total visit is

$$\sum_{n=1}^{\infty} n f_{ii}^{n-1}(1-f_{ii}) = \frac{1}{1-f_{ii}}.$$

12.1.5 Binomial distribution

Definition 12.1.4 (binomial distribution). *A discrete random variable X is said to have a Binomial distribution $Binomial(n,p)$ with parameter n,p if it has pmf given as*

$$f(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$$

with $x \in \{0,1,2,...,n\}$.

Remark 12.1.1 (interpretation). Binomial distribution represents the probability distribution of the number of successes in a sequence of n independent binary experiments, each of which yields 1 with probability p.

Remark 12.1.2 (relation to Bernoulli distribution). Let X_i be iid random variables with Bernoulli distribution of parameter p, then

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

is a random variable of binomial distribution with parameter (n,p).

Lemma 12.1.6 (sum of independent binomial random variable). *Let $X_1,X_2,...,X_K$ be the independent binomial random variables with parameter $(n_1,p),(n_2,p),...,(n_K,p)$. Let $Y=\sum_{i=1}^{K}X_i$. Then*

- $M_{X_i}(t)=(1-p+pe^t)^{n_i}, i=1,...,K.$
- $M_Y(t)=(1-p+pe^t)^{\sum_{i=1}^{K}n_i}$
- $Y \sim Binomial(\sum_{i=1}^{K}n_i,p)$.

Proof. (1) Use the mgf of Bernoulli distribution [Lemma 12.1.1]. (2)(3) Consider $X_1 \sim Binomial(n_1,p)$ and $X_2 \sim Binomial(n_2,p)$, each has momemt generating function of $(1-p+pe^t)^{n_1}$ and $(1-p+pe^t)^{n_2}$. X_1+X_2 will have mgf of $(1-p+pe^t)^{n_1+n_2}$ [Theorem 11.6.2], corresponding to $Binomial(n_1+n_2,p)$. It is straight forward to extend multiple cases. □

Lemma 12.1.7 (convergence of binomial distribution to Poisson distribution). *Suppose that $p_n \in (0,1)$ for $n \in \mathcal{N}_+$ and $np_n \to \lambda$ as $n \to \infty$. Then the binomial distribution*

with parameters n and p_n converges to the Poisson distribution with parameter λ ***in distribution*** *as $n \to \infty$. That is, for fixed $k \in \mathcal{N}$,*

$$\binom{n}{k} p_n^k (1-p_n)^{n-k} \to e^{-\lambda} \frac{\lambda^k}{k!}$$

as $n \to \infty$.

Proof. (direct method)Note that

$$\begin{aligned}
\binom{n}{k} p_n^k (1-p_n)^{n-k} &= \frac{n(n-1)(n-2)\cdots(n-k+1)}{k!} (p_n)^k (1-p_n)^{n-k} \\
&= \frac{n(n-1)(n-2)\cdots(n-k+1)}{k!} (\frac{\lambda}{n})^k (1-\frac{\lambda}{n})^{n-k} \\
&\approx \frac{n^k}{k!} (\frac{\lambda}{n})^k (1-\frac{\lambda}{n})^{n-k} \\
&= \frac{\lambda^k}{k!} (1-\frac{\lambda}{n})^{n-k} \\
&\to e^{-\lambda \frac{n-k}{n}} \frac{\lambda^k}{k!} \\
&\approx e^{-\lambda} \frac{\lambda^k}{k!}
\end{aligned}$$

(use generating function) Note that binomial distribution has probability generating function **[??]**

$$((1-p_n)+p_n s)^n = (1+(pns-p_n)n/n)^n \to e^{(s-1)a}, n \to \infty$$

where $e^{(s-1)a}$ is the generating function of Poisson distribution. □

Remark 12.1.3 (Poisson distribution as an approximate for large n and small k)**.** Note that the lemma requires that k fixed. In other words, when $n \gg k$, we can use Poisson distribution to approximate binomial distribution.

12.1.6 Normal distribution

Definition 12.1.5 (normal distribution). *A random variable X with normal distribution $N(\mu, \sigma^2)$, characterized by parameters μ and σ, has its pdf given by*

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

X is called normal random variable, or Gaussian random variable. If $\mu = 0, \sigma = 1$, X is also called standard normal random variable.

Lemma 12.1.8 (moment generating function). *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

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

Proof. (1) $m_X(t) = E[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$ complete the square and get the result. (2)Let $Y = \sigma X + \mu$ and use Theorem 11.6.2. Then $m_Y(t) = e^{\mu t} m_X(\sigma t)$ □

Lemma 12.1.9 (basic properties of normal random variable). *Consider $X \sim N(\mu_x, \sigma_x^2), Y \sim N(\mu_y, \sigma_y^2)$.*

- *If X, Y **are independent**, then we have*
  - $$aX + b \sim N(a\mu + b, a^2\sigma_x^2)$$
  - $$X + Y \sim N(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$$
    $$aX + bY \sim N(a\mu_x + b\mu_y, a^2\sigma_x^2 + b^2\sigma_y^2)$$
- *If X **and** Y **are not independent but jointly normal**, then $X + Z$ will be normal, and*
  $$aX + bZ \sim N(a\mu_x + b\mu_y, a^2\sigma_x^2 + b^2\sigma_z^2 + 2abCov(X, Z)).$$
- *Assume X, Y are independent. Further let $W = \rho X + \sqrt{1-\rho^2}Y, \rho \in [-1, 1]$. Then W is normal, correlated with X and Y, and the sum $X + W$ is also normal; that is,*
  $$W \sim N, aX + bW \sim N.$$
- ***In general, the sum of two dependent normal random variable is not necessarily normal.*** *See Corollary 12.1.1.1.*

Proof. (1) Directly from the properties of moment generating functions at Theorem 11.6.2. (2)The proof of two general jointly normal random variable will be showed in Corollary 12.1.1.1. (3) Note that

$$(X, W)^T = (X, \rho X + \sqrt{1-\rho^2}Y)^T = \begin{bmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{bmatrix} (X, Y)^T,$$

therefore (X, W) are jointly normal [Theorem 14.1.1]. Then we use (2). □

Lemma 12.1.10 (moments of standard normal distribution). *Let $X \sim N(0,1)$, then*

$$E[X] = 0, E[X^2] = 1, E[X^3] = 0, E[X^4] = 3$$

Moreover, all odd moments are 0.

Proof. The mgf is $m(t) = e^{t^2/2}$, then

$$m'(t) = te^{t^2/2}$$
$$m''(t) = e^{t^2/2} + t^2 e^{t^2/2}$$
$$\dots$$

For all odd moments

$$\int x^{2k+1} f(x) dx$$

has integrand as odd function. □

Corollary 12.1.0.1 (moments of normal distribution). *Let $X \sim N(0, \sigma^2)$, then*

$$E[X] = 0, E[X^2] = \sigma^2, E[X^3] = 0, E[X^4] = 3\sigma^4$$

Moreover, all odd moments are 0.

Proof. The mgf is $m(t) = e^{\sigma^2 t^2/2}$, then

$$m'(t) = \sigma^2 te^{\sigma^2 t^2/2}$$
$$m''(t) = \sigma^2 e^{\sigma^2 t^2/2} + \sigma^4 t^2 e^{t^2/2}$$
$$\dots$$

For all odd moments

$$\int x^{2k+1} f(x) dx$$

has integrand as odd function. □

12.1.7 Half-normal distribution

Definition 12.1.6 (half-normal distribution). *Let X follow an ordinary normal distribution $N(0,\sigma^2)$. Then $Y=|X|$ follows a **half-normal distribution** with parameter σ. It has probability density function*

$$f_Y(y;\sigma)=\frac{\sqrt{2}}{\sigma\sqrt{\pi}}\exp(-\frac{y^2}{2\sigma^2}).$$

Lemma 12.1.11 (basic properties of half normal distribution). *Let Y follow a half-normal distribution with parameter σ. Then*

- $E[Y]=\frac{\sigma\sqrt{2}}{\sqrt{\pi}}$.
- $Var[Y]=\sigma^2(1-\frac{2}{\pi})$

12.1.8 Laplace distribution

Definition 12.1.7 (Laplace distribution). *A random variable X has a **Laplace distribution**, denoted by $Lap(\mu,b)$, if its probability density function is*

$$f(x|\mu,b)=\frac{1}{2b}\exp(-\frac{|x-\mu|}{b})=\begin{cases}\frac{1}{2b}\exp(-\frac{\mu-x}{b}),if\ x<\mu\\ \frac{1}{2b}\exp(-\frac{x-\mu}{b}),if\ x\geq\mu\end{cases}.$$

Lemma 12.1.12 (properties of Laplace distribution). *Let X be a random variable with Laplace distribution with parameter μ,b. It follows that*

- *The mean and the median are μ.*
- *The variance is $2b^2$.*
- *The cdf is given by*

$$\begin{aligned}F(x)=\int_{-\infty}^{x}f(u)du&=\begin{cases}\frac{1}{2}\exp(-\frac{\mu-x}{b}),if\ x<\mu\\ 1-\frac{1}{2}\exp(-\frac{x-\mu}{b}),if\ x\geq\mu\end{cases}\\ &=\frac{1}{2}+\frac{1}{2}sgn(x-\mu)(1-\exp(-\frac{|x-\mu|}{b})).\end{aligned}$$

- *The inverse cdf is given by*

$$F^{-1}(p) = \mu - b \cdot sgn(p - 0.5) \ln(1 - 2|p - 0.5|).$$

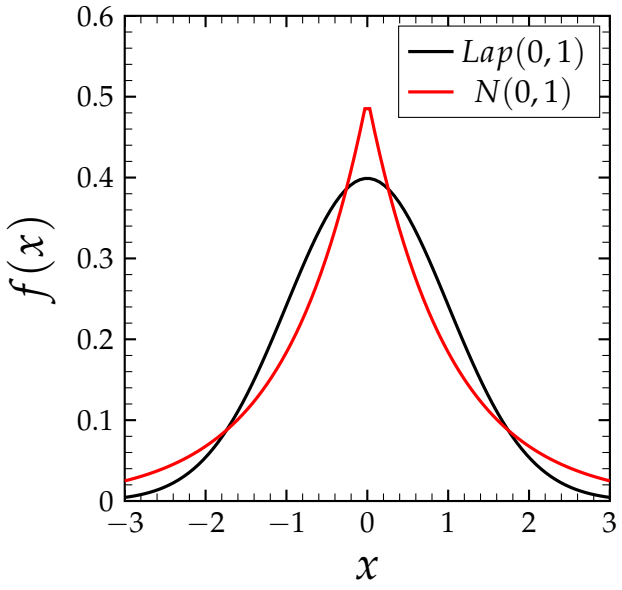


Figure 12.1.1: Comparison of Laplace distribution and normal distribution.

12.1.9 Multivariate Gaussian/normal distribution

12.1.9.1 *Basic definitions*

Definition 12.1.8 (multivariate Gaussian/normal distribution). *A random vector is said be multivariate Gaussian/normal random variable if it's pdf is multivariate Gaussian/normal distribution, whose support is $\mathbb{R}^n$ and its pdf is*

$$\rho(x; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|det\Sigma|^{1/2}} \exp(-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu))$$

with mean $\mu \in \mathbb{R}^n$ and covariance matrix $\Sigma \in \mathbb{R}^{n \times n}$.

Lemma 12.1.13 (mgf of multivariate Gaussian/normal random variables). *[3, p. 181] An n-dimensional random vector $X \sim MN(\mu, \Sigma)$ has its mgf given by*

$$M_X(t) \triangleq E[\exp(t^T X)] = \exp(t^T \mu + \frac{1}{2} t^T \Sigma t)$$

for all $t \in \mathbb{R}^n$.

Proof.

$$\begin{aligned} M_X(t) &= E[\exp(t^T X)] \\ &= \exp(E[t^T X] + \frac{1}{2} Var[t^T X]) \\ &= \exp(t^T \mu + \frac{1}{2} t^T \Sigma t) \end{aligned}$$

where we use the fact that $t^T X \sim N(t^T \mu, t^T \Sigma t)$ from Theorem 14.1.1. □

Remark 12.1.4 (implication). Given a random vector X, if we want to check whether X is a multivariate Gaussian, we can check its mgf. If its mgf is the exponential of a linear form plus a quadratic form, then it is multivariate Gaussian.

Lemma 12.1.14 (criterion via linear combination). *A vector $X = (X_1, X_2, ..., X_n)^T$ is a multivariate Gaussian distribution if every linear combination*

$$S = a^T X, a \in \mathbb{R}^n$$

has a normal distribution.

Proof. Because $a^T X$ is normal, then it has characteristic function

$$E[\exp(ita^T X)] = \exp(ita^T \mu_X - \frac{1}{2} t^2 (a)^T \Sigma_X a).$$

Since a is arbitrary, we can say for any $t' \in \mathbb{R}^n$, we have

$$E[\exp(it' X)] = \exp(i[t']^T \mu_X - \frac{1}{2} [t']^T \Sigma_X t).$$

That is, X is multivariate Gaussian. □

Example 12.1.5 (bivariate Gaussian distribution). Let $f(x,y)$ be the density of a bivariate Gaussian distribution $MN(\mu,\Sigma)$, where

$$\mu = \begin{Bmatrix} \mu_X \\ \mu_Y \end{Bmatrix}, \Sigma = \begin{Bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{Bmatrix}.$$

Then,

$$f(x,y) = \frac{\exp(-\frac{1}{2(1-\rho^2)}[\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - 2\frac{\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}])}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}.$$

12.1.9.2 *Affine transformation and its consequences*

Theorem 12.1.1 (affine transformation for multivariate normal distribution). *Let X be an n-dimensional random vector with $MN(\mu,\Sigma)$ distribution. Let $Y = AX + b, A \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^m$. Then Y is an m-dimensional random vector having a $MN(A\mu + b, A\Sigma A^T)$ distribution.*

Proof. Use moment generating function to prove. Let $Y = AX + b$, then from Lemma 11.6.5

$$M_Y(t) = e^{t^T b} M_X(A^T t) = e^{t^T(A\mu+b)+\frac{1}{2}t^T A\Sigma A^T t}$$

which indicates $Y \sim MN(A\mu + b, A\Sigma A^T)$. □

Corollary 12.1.1.1 (sum of two multivariate normal random vectors). *Let $X_1 \sim MN(\mu_1,\Sigma_1)$ and $X_2 \sim MN(\mu_1,\Sigma_1)$ be two n dimensional multivariate normal random variable. It follows that*

- *If X_1 and X_2 are independent, then $Y = X_1 + X_2$ is a multivariate normal random vector with $MN(\mu_1 + \mu_2, \Sigma_1 + \Sigma_2)$.*
- *If X_1 and X_2 are dependent and (X_1, X_2) are **jointly normal** [a] with covariance matrix given by*

$$\Sigma = \begin{bmatrix} \Sigma_1 & \Sigma_{12} \\ \Sigma_{12} & \Sigma_2 \end{bmatrix},$$

then $Y = X_1 + X_2$ is a multivariate normal random vector with $MN(\mu_1 + \mu_2, \Sigma_1 + \Sigma_2 + 2\Sigma_{12})$.

[a] If X_1 and X_2 are not jointly normal, Y is not normal.

Proof. (1) Consider a $2n$-dimensional multivariate normal random variable Z with distribution $\mu = [\mu_1; \mu_2]$, $\Sigma = \Sigma_1 \oplus \Sigma_2$ (Σ is a diagonal block matrix with two blocks Σ_1 and Σ_2). Then we can construct a linear transformation matrix

$$A = \begin{bmatrix} I_n & I_n \end{bmatrix}$$

to construct $Y = AZ$. Apply affine transformation theorem [Theorem 14.1.1] to $Y = AZ$, we have $Y \sim MN(\mu_1 + \mu_2, \Sigma_1 + \Sigma_2)$. (2) same as (1). □

Note 12.1.1. caution! The joint distribution of two Gaussian margins are not necessarily joint Gaussian:

- Two multivariate normal random variables are not necessarily joint normal.[a]. For example, consider two marginal distribution of Gaussian. For Gaussian copula, the joint distribution is multivariate Gaussian; however, for other copulas including Frank copula and Clayton copula, the joint distribution is not multivariate Gaussian.
- If two multivariate normal random variables are independent, then they are joint normal.

[a] link

Corollary 12.1.1.2 (orthonormal transformation maintains independence). *Let X be a n dimensional random vector with $MN(0, I)$. If C is an orthonormal matrix, then $Y = CX$ has distribution $MN(0, I)$. That is, orthonormal transformation will preserve independence.*

Proof. $Cov(Y) = C^T I C = I$. □

12.1.9.3 *Marginal and conditional distribution*

Lemma 12.1.15 (marginal distribution). *The multivariate Gaussian distribution $\rho(x;\mu,\Sigma)$ on $\mathbb{R}^n$ has marginal distribution on $\mathbb{R}^k, k \leq n$ given as $\rho(x_1;\mu_1,\Sigma_{11}), x_1 \in \mathbb{R}^k$ where we decompose*

$$\mu = [\mu_1,\mu_2]^T, \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

Proof. Use above Theorem 14.1.1. Let

$$A = \begin{bmatrix} I & 0 \end{bmatrix}.$$

Then $X_1 = AX$. □

Lemma 12.1.16 (full joint distribution can be constructed from pair joint distribution). *Let $X = (X_1, X_2, ..., X_n)^T$ be a random multivariate Gaussian vector with mean $\mu = (\mu_1, \mu_2, ..., \mu_n)^T$ and covariance matrix $\Sigma \in \mathbb{R}^{n\times n}$. The the pair $(X_i, X_j), i \neq j$ has joint distribution*

$$\hat{\mu} = (\mu_i, \mu_j), \hat{\Sigma} \in \mathbb{R}^{2\times 2}, \hat{\Sigma}_{11} = \Sigma_{ii}, \hat{\Sigma}_{12} = \Sigma_{ij}.$$

That is, all the pair joint distribution can construct the full joint distribution.

Proof. Directly from Lemma 14.1.3. □

Remark 12.1.5 (caution! not all the distribution has this property). If the full joint distribution is not Gaussian, then such property (reconstruct full distribution from pair distribution) will not generally hold.

Theorem 12.1.2 (conditional distribution). *The multivariate Gaussian distribution $\rho(x;\mu,\Sigma)$ on $\mathbb{R}^n$ has a conditional Gaussian distribution on $\mathbb{R}^k, k \leq n$ given by*

$$\frac{f(x_1,x_2)}{f(x_2)} = MN(x_1;\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$$

where we decompose

$$\mu = [\mu_1^T,\mu_2^T]^T, \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

with $\mu_1 \in \mathbb{R}^k, \mu_2 \in \mathbb{R}^{n-k}$.

Proof. See link □

Remark 12.1.6 (gaining information). From the conditional distribution, we can see that given the information of x_2, the mean of x_1 will be corrected and the variance of x_1 will be reduced.

Example 12.1.6 (bivariate Gaussian distribution). Let $f(x,y)$ be the density of a bivariate Gaussian distribution $MN(\mu,\Sigma)$, where

$$\mu = \begin{Bmatrix} \mu_X \\ \mu_Y \end{Bmatrix}, \Sigma = \begin{Bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{Bmatrix}.$$

Then,

$$X|Y \sim N(\mu_X + \frac{\rho\sigma_X}{\sigma_Y}(y-\mu_Y), (1-\rho^2)\sigma_X^2).$$

12.1.9.4 *Box Muller transformation*

Lemma 12.1.17 (Box Muller transformation). *Let $X, Y \sim N(0,1)$ and X, Y be independent. Let*

$$R = \sqrt{X^2+Y^2}, \Theta = \arctan(Y/X)$$

. Then

- *R and Θ are independent.*
- *$\Theta \sim U(0, 2\pi)$ and $F_R(r) = 1 - \exp(-r^2/2)$.*
- *Suppose we have U_1, U_2 being independent uniform on $[0,1]$. Then $2\pi U_1$ and $\sqrt{-2\ln(1-U_2)}$ are independent and have the same distribution of R and Θ.*
- *Further, $\sqrt{-2\ln(1-U_2)}\cos(2\pi U_1)$ and $\sqrt{-2\ln(1-U_2)}\sin(2\pi U_1)$ are independent and have the same distribution of X and Y.*

Proof. (1)Using polar transformation Lemma 11.4.9, we have

$$\begin{aligned} Pr(R<r, \Theta<\theta) &= \int_0^r \int_0^\theta \frac{1}{2\pi}\exp(-\frac{r^2}{2}) r dr d\theta \\ &= \int_0^r \int_0^\theta \exp(-\frac{r^2}{2}) r dr \frac{1}{2\pi} d\theta \\ &= F_R(R<r) F_\Theta(\Theta<\theta) \end{aligned}$$

Using independence condition Lemma 11.4.3, we know that R and Θ are independent. (2) Integrate directly in (1). (3) Let $U = 1 - \exp(-R^2/2)$. Based on probability integral

transform [Lemma 14.4.2], we know U is an uniform random variable. Or equivalently, $R = \sqrt{-2\ln(1-U)}$ has the same distribution of R. (4) Note that $X = R\cos(\Theta), Y = R\sin(\Theta)$. □

12.1.10 Lognormal distribution

12.1.10.1 *Univariate lognormal distribution*

Definition 12.1.9 (lognormal distribution). *A random variable Y has a lognormal distribution with parameters μ and σ^2, written as*

$$Y \sim LN(\mu, \sigma^2)$$

if $\log(Y)$ is normally distributed as $N(0, \sigma^2)$. Several equivalent definitions are:

- *$Y \sim LN(\mu, \sigma^2)$ if and only if $\log(Y) \sim N(\mu, \sigma^2)$.*
- *$Y \sim LN(\mu, \sigma^2)$ if and only if $Y = e^X$ with $X \sim N(\mu, \sigma^2)$.*
- *The distribution function is given as*

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

Lemma 12.1.18 (basic properties of lognormal distribution). *Let $Y \sim LN(\mu, \sigma^2)$, or equivalently $Y = \exp(X), X \sim N(\mu, \sigma^2)$ then*

- *The distribution function for Y is given as*

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

- $E[Y] = \exp(E[X] + \frac{1}{2}Var[X^2]) = \exp(\mu + \sigma^2/2)$.
- $E[Y^2] = \exp(2\mu + 2\sigma^2), E[Y^m] = \exp(m\mu + \frac{1}{2}m^2\sigma^2)$
- *$Var[Y] = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1)$. In particular $\mu = 0$, we have*

$$E[Y] = \exp(\frac{1}{2}\sigma^2), E[Y^m] = \exp(\frac{1}{2}m^2\sigma^2), Var[Y] = \exp(2\sigma^2) - \exp(\sigma^2).$$

- *If $X_1 \in N(\mu_1, \sigma_1^2), X_2 \in N(\mu_2, \sigma_2^2)$, then*

$$E[\exp(X_1 + X_2)] = \exp(E[X_1] + E[X_2] + \frac{1}{2}Var[X] + \frac{1}{2}Var[X_2] + Cov(X_1, X_2)).$$

- $$\mu = \log(\frac{E[Y]^2}{\sqrt{E[Y^2]}}), \sigma^2 = \ln(\frac{E[Y^2]}{E[Y]^2}).$$
- *The median of* Y *is* $\exp(\mu)$.
- *skewness*
$$(\exp(\sigma^2)+2)\sqrt{\exp(\sigma^2)-1} > 0.$$

Proof. (1) Note that

$$x = \ln y, f_Y(y) = f_X(\ln y)\left|\frac{d\ln y}{dy}\right|.$$

(2)(3) Note that for $X \sim N(\mu, \sigma^2)$, $M_X(t) = \exp(\mu t + \frac{1}{2}\sigma^2 t^2)$. Then

$$E[Y] = E[\exp(X)] = M_X(1) = \exp(\mu + \frac{1}{2}\sigma^2),$$

and

$$E[Y^2] = E[\exp(2X)] = M_X(2) = \exp(2\mu + 2\sigma^2),$$

and

$$Var[Y] = E[Y^2] - (E[Y])^2 = \exp(2\mu + 2\sigma^2) - \exp(2\mu + \sigma^2).$$

(4) Note that $X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2)$. Then we use (1). (5) Note that the exponential is a monotone function, the median of Y will be $\exp(median\ X) = \exp(\mu)$, where we used the fact that median of X is μ. □

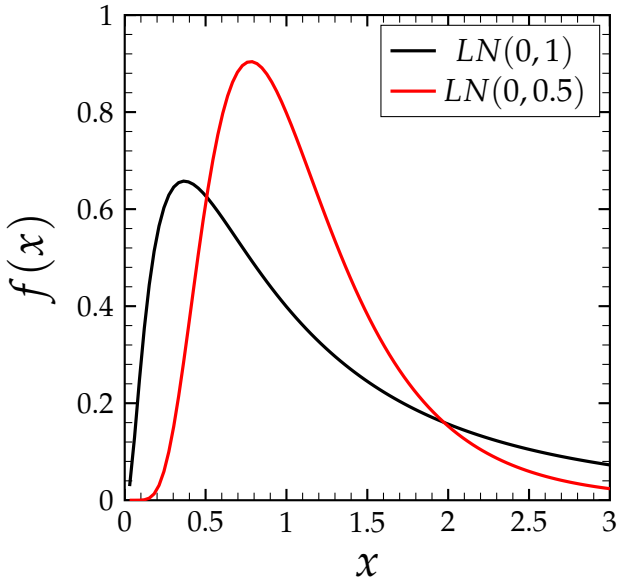


Figure 12.1.2: Density of $LN(0,1)$ and $LN(0,0.5)$. Note the positive skewness.

12.1.10.2 *Extension to univariate lognormal distribution*

Definition 12.1.10. *[4]*

- ***regular log-normal distribution*** *with parameter* (μ, σ^2) *is given by*

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp(-\frac{(\ln x - \mu)^2}{2\sigma^2}), x > 0.$$

- ***negative log-normal distribution*** *with parameter* (μ, σ)*, denoted by* $NLN(\mu, \sigma^2)$ *is given by*

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp(-\frac{(\ln -x - \mu)^2}{2\sigma^2}), x < 0.$$

- ***shifted log-normal distribution*** *with parameter* (μ, σ, τ)*,denoted by* $SLN(\mu, \sigma^2, \tau)$ *is given by*

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

- ***negative shifted log-normal distribution*** *with parameter* (μ, σ^2, τ) *is given by*

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

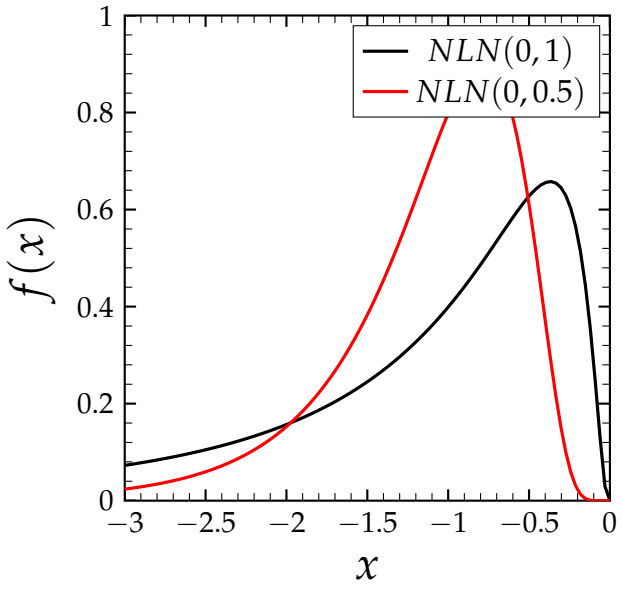


Figure 12.1.3: Density of $NLN(0,1)$ and $NLN(0,0.5)$. Note the negative skewness.

Lemma 12.1.19. *Let $X \sim LN(\mu, \sigma^2)$. It follows that*

- *Let $Y = -X$. Then $Y \sim NLN(\mu, \sigma^2)$.*
- *Let $Z = X + \tau$. Then $Y \sim NLN(\mu, \sigma^2, \tau)$.*
- *Let $W = -X - \tau$. Then $Y \sim NSLN(\mu, \sigma^2, \tau)$.*

Proof. Straight forward from definition and transformation. □

Lemma 12.1.20 (basic properties of shifted lognormal distribution). *Let $X \sim SLN(\mu, \sigma^2, \tau)$. Then*

- $$E[X] = \tau + \exp(\mu + \frac{1}{2}\sigma^2).$$
- $$E[X^2] = \tau^2 + 2\tau \exp(\mu + \frac{1}{2}\sigma^2) + \exp(2\mu + 2\sigma^2).$$
- $$E[X^3] = \tau^3 + 3\tau^2 \exp(\mu + \frac{1}{2}\sigma^2) + 3\tau \exp(2\mu + 2\sigma^2) + \exp(3\mu + \frac{9}{2}\sigma^2).$$

Proof. Note that from Lemma 12.1.18, we have if $Y \sim LN(\mu, \sigma^2)$, then $E[Y^m] = \exp(m\mu + \frac{1}{2}m^2\sigma^2)$. Then, we use

$$E[X] = E[Y + \tau] = E[Y] + \tau,$$
$$E[X^2] = E[(Y + \tau)^2] = E[Y^2] + 2\tau E[Y] + \tau^2,$$
$$E[X^3] = E[(Y + \tau)^3] = E[Y^3] + 3\tau E[Y^2] + 3\tau^2 E[Y] + \tau^3.$$

□

12.1.10.3 *Multivariate lognormal distribution*

Definition 12.1.11 (multivariate lognormal distribution). *If $X = (X_1, X_2, ..., X_n) \sim MN(\mu, \Sigma)$, then $Y = \exp(X) = (\exp(X_1), \exp(X_2), ..., \exp(X_n)) \sim MLN(\mu, \Sigma)$, i.e., Y has multivariate lognormal distribution*

Lemma 12.1.21 (basic properties of multivariate lognormal distribution). *Let $X = (X_1, X_2, ..., X_n) \sim MN(\mu, \Sigma)$ and $Y = \exp(X) = (\exp(X_1), \exp(X_2), ..., \exp(X_n))$. Then*

- $E[Y_i] = \exp(\mu_i + \frac{1}{2}\Sigma_{ii})$.

- $E[Y_iY_j] = \exp(\mu_i + \mu_j + \frac{1}{2}(\Sigma_{ii} + \Sigma_{jj} + 2\Sigma_{ij})) = E[Y_i][Y_i]\exp(\Sigma_{ij})$.
- $Var[Y_i] = \exp(2\mu_i + \Sigma_{ii})(\exp(\Sigma_{ii}) - 1)$.
- $Cov[Y_iY_j] = \exp(\mu_i + \mu_j + \frac{1}{2}(\Sigma_{ii} + \Sigma_{jj}))(\exp(\Sigma_{ij}) - 1)$.

Proof. (1)Note that $M_X(t) = \exp(t^T\mu + \frac{1}{2}t^T\Sigma t), t \in \mathbb{R}^n$, and $E[Y_i] = M_X(e_i)$. (2) Let $t = e_i + e_j$. Then

$$E[Y_iY_j] = M_X(t) = \exp(\mu_i + \mu_j + \frac{1}{2}(\Sigma_{ii} + \Sigma_{jj} + 2\Sigma_{ij})).$$

(3)

$$Var[Y_i] = E[Y_iY_i] - E[Y_i]E[Y_i].$$

(4)

$$Cov[Y_i, Y_j] = E[Y_iY_j] - E[Y_i]E[Y_j].$$

□

12.1.11 Exponential distribution

Definition 12.1.12 (exponential distribution). *A random variable X is said to have an exponential distribution $Exp(\lambda)$ with parameter λ if it has pdf given as*

$$p(x|\lambda) = \lambda \exp(-\lambda x)$$

with $x \in [0, \infty)$.

Lemma 12.1.22 (basic properites). *Let X be a random variable with exponential distribution with parameter λ, then we have*

- $E[X] = 1/\lambda$
- $Var[X] = 1/\lambda^2$
- *memoryless:*

$$P(X > s + t|X > s) = P(X > t)$$

(even though $P(X > s + t) < P(X > t)$)

Proof. (1)(2) are straightforward. (3) The cmf is given as

$$F(t) = \int_0^t \lambda \exp(-\lambda\tau)d\tau = 1 - \exp(-\lambda t)$$

$$P(X > s+t|X > s) = \frac{P(X > s+t \cap X > s)}{P(X > s)} = \frac{P(X > s+t)}{P(X > s)} = \frac{\exp(-\lambda(s+t))}{\exp(-\lambda t)} = \exp(-\lambda t)$$

□

Remark 12.1.7 (interpretation of memorylessness). Supppose we are waiting for an event to occur, and we model the waiting time as a random variable X with $Exp(\lambda)$. If we already wait for s time, the distribution that we need to wait an extra of t time is the same as the distribution of the waiting time at time 0. **Exponential distribution is the only memoryless continuous distribution**[5].

Lemma 12.1.23 (Normal approximate sum of Exponential). *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 $\sigma = n/\lambda^2$.

Proof. Directly from Central Limit Theorem [Theorem 11.11.3]. Also see Gamma distribution properties, since exponential distribution is a special case of Gamma distribution. □

12.1.12 Gamma distribution

Definition 12.1.13 (Gamma distribution). *[6, p. 42] A random variable X is said to have a Gamma distribution $Gamma(a,b)$ with parameter a, b if it has pdf given as*

$$f(x) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}$$

with support $x \in (0, \infty)$.

Remark 12.1.8 (exponential distribution is a special case). An exponential distribution with parameter b is a Gamma distribution $Gamma(1,b)$with

$$f(x) = be^{-bx}.$$

Remark 12.1.9 (Application in arrival times of Poisson process). If $N(t)$ is a Poisson process with rate λ, then the arrival time $T_1, T_2, ...$ have $T_n \sim Gamma(n, \lambda)$ distribution.(See Lemma 18.5.4)

Caution! Gamma distribution is different from Gamma function $\Gamma(t)$, which is given as

$$\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} dx$$

Remark 12.1.10 (conjugate prior for Poisson distribution). Gammad distribution conjugate prior for the parameter of Poisson distribution. When integrate out x in $\Gamma(t)$, we have

$$\int_0^\infty x^{a-1} e^{-bx} dx = \Gamma(a)/b^2$$

Lemma 12.1.24 (mean and variance). *The Gamma distribution $Gamma(a, b)$ has mean a/b and variance a/b^2.*

Proof. Using the property of

$$\int_0^\infty x^{a-1} e^{-bx} dx = \Gamma(a)/b^a,$$

we can show the result. □

Theorem 12.1.3 (sum of Gamma random variables). *[3, p. 163] Let $X_1, ..., X_n$ be independent random variables. Suppose $X_i \sim Gamma(a_i, b), \forall i = 1, ..., n$. Then*

$$Y = \sum_{i=1}^n X_i \sim Gamma(\sum_{i=1}^n a_i, b)$$

Proof. This can be proved using moment generating functions. □

Lemma 12.1.25 (Normal approximate sum of Gamma). *Let $X_1, ..., X_n$ be independent iid random variables of $Gamma(a, b)$, 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 $\mu = na/b$,and $\sigma = na/b^2$.

Proof. Directly from Central Limit Theorem [Theorem 11.11.3]. □

12.1.13 Hypergeometric distribution

Definition 12.1.14 (hypergeometric distribution distribution). *[3, p. 148] A random variable X is said to have a hypergeometric distribution $HG(N,K,n)$ with parameter N,K,n if it has pmf given as*

$$p(x=k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

with support $x \in \{0,1,...,\min(n,K)\}$. Note that the parameters should be non-negative integers and satisfying

$$N \geq K, N \geq n.$$

Remark 12.1.11 (interpretation). $p(x=k)$ describes the probability of k successes in n draws, without replacement, from a finite population of size N that contains exactly K successes.

Lemma 12.1.26 (combinatorial identities). *Assuming $K \geq n$, we have*

$$\sum_{0 \leq k \leq n} \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} = 1$$

Lemma 12.1.27 (mean of a hypergeometric distribution). *[3, p. 148] Let X be a random variable with $HG(N,K,n)$, then its mean is*

$$E[X] = n\frac{K}{N}$$

12.1.14 Beta distribution

Definition 12.1.15 (Beta distribution). *A random variable X is said to have a Beta distribution $B(a,b)$ with parameter a,b if it has a pdf given as*

$$f(x)=\frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}, B(a,b)=\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

with support $x\in[0,1]$.

Remark 12.1.12.

- $$\int_0^1 x^{a-1}(1-x)^{b-1}dx=\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$
- Beta distribution is commonly **used as the conjugate prior for binomial distribution**, where
  $$p(y_1,...,y_n|\theta)=\theta^{\sum_i y_i}(1-\theta)^{n-\sum_i y_i}, y_i\in\{0,1\}$$
  then the posterior distribution will also be Beta.

Lemma 12.1.28 (basic property). *Let X be a random variable with distribution $B(a,b)$.*

- $$E[X]=\frac{a}{a+b}.$$
- $$E[X^2]=\frac{a(a+1)}{(a+b)(a+b+1)}.$$
- $$E[X^r]=\frac{a(a+1)\cdots(a+r-1)}{(a+b)(a+b+1)\cdots(a+b+r-1)}.$$
- $$Var[X]=\frac{ab}{(a+b)^2(a+b+1)}.$$
- *The mode of X, i.e., the value x that has the maximum probability is*
  $$x^*=\frac{a-1}{a+b-2}.$$

Proof. (1)This can be proved using properties of Gamma distribution.

$$\begin{aligned}
E[X] &= \int_0^1 xf(x)dx \\
&= \int_0^1 \frac{x^a(1-x)^{b-1}}{B(a,b)} \\
&= \frac{B(a+1,b)}{B(a,b)} \\
&= \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} \Big/ \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \\
&= \frac{\Gamma(a+1)\Gamma(a+b)}{\Gamma(a+b+1)\Gamma(a)} \\
&= \frac{a}{a+b}
\end{aligned}$$

(2)

$$\begin{aligned}
E[X^2] &= \int_0^1 x^2 f(x)dx \\
&= \int_0^1 \frac{x^{a+1}(1-x)^{b-1}}{B(a,b)} \\
&= \frac{B(a+2,b)}{B(a,b)} \\
&= \frac{\Gamma(a+2)\Gamma(b)}{\Gamma(a+b+2)} \Big/ \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \\
&= \frac{\Gamma(a+2)\Gamma(a+b)}{\Gamma(a+b+2)\Gamma(a)} \\
&= \frac{a(a+1)}{(a+b)(a+b+1)}
\end{aligned}$$

(3) Use $Var[X] = E[X^2] - E[X]^2$. (4) To find the maximizer for $x^{a-1}(1-x)^{b-1}$, we take the log and maximize it. We have

$$\ln f(x) = (a-1)\ln x + (b-1)\ln(1-x).$$

Take the derivative with respect to x and set to 0, we have

$$\begin{aligned}
\frac{a-1}{x} &= \frac{b-1}{1-x} \\
(a-1)(1-x) &= x(b-1) \\
\implies x^* &= \frac{a-1}{a+b-2}.
\end{aligned}$$

□

12.1.15 Multinomial distribution

Definition 12.1.16. *[6, p. 35] A discrete random vector $X = (X_1, ..., X_n)$ is said to have multinomial distribution with parameters $(p_1, ..., p_n)$ and m if its pmf is given as*

$$f(x_1, x_2, ..., x_n) = \frac{m!}{x_1!...x_n!} p_1^{x_1} ... p_n^{x_n}$$

where we require $x_i \in \{0, ..., m\}, \sum x_i = m, \sum p_i = 1$.

Remark 12.1.13. Consider m independent experiment, each has n outcomes with probability p_i to occur. The outcome distribution is given as[7]

$$f(x_1, x_2, ..., x_n) = \frac{m!}{x_1!...x_n!} p_1^{x_1} ... p_n^{x_n}$$

where $\sum x_i = m, \sum p_i = 1$.

Lemma 12.1.29 (basic properties of Multinomial Distribution). *Let $X = (X_1, ..., X_n)$ discrete random vector with multinomial distribution with parameters $p = (p_1, ..., p_n)$ and m.*

-
- $$E[X_i] = np_i.$$
- $$Var[X_i] = np_i(1 - p_i), Cov(X_i, X_j) = np_i(1 - p_i),$$
  or in vector form
  $$Var[X] = n(diag(p) = pp^T).$$

Proof. (1)This can be proved using properties of Gamma distribution.

$$\begin{aligned} E[X_i] &= \int_0^1 x_i f(x) dx \\ &= \frac{\prod_{k=1}^K \Gamma(a_k + \delta_{ik})}{\Gamma(a_0 + 1)} / \frac{\prod_{k=1}^K \Gamma(a_k)}{\Gamma(a_0)} \\ &= \frac{a_i}{a_0} \end{aligned}$$

(2)

$$\begin{aligned} E[X_i^2] &= \int_0^1 x_i^2 f(x) dx \\ &= \frac{\prod_{k=1}^K \Gamma(a_k + 2\delta_{ik})}{\Gamma(a_0 + 2)} / \frac{\prod_{k=1}^K \Gamma(a_k)}{\Gamma(a_0)} \\ &= \frac{a_i(a_i+1)}{(a_0+1)a_0} \end{aligned}$$

(3) Use $Var[X] = E[X^2] - E[X]^2$. (4) To find the maximizer for $f(x)$, we take the log and maximize it. The optimality condition requires that $x_i^* \propto a_i - 1$ and $\sum_{i=1}^K a_i = 1$. □

12.1.16 Dirichlet distribution

Definition 12.1.17. *[6, p. 49] A random vector $X = (X_1, ..., X_K)$ is said to have a Dirichlet distribution distribution with parameter $a = (a_1, ..., a_K)$ if it has pdf given as*

$$f(x_1, ..., x_K) = \frac{1}{B(a)} \prod_{k=1}^K x_k^{a_k - 1}$$

with support $x \in \{x : 0 \leq x_k \leq 1, \sum_k x_k = 1, \forall k = 1, 2..., K\}$, and $B(a)$ is a normalization constant given as

$$B(a) = \frac{\prod_{k=1}^K \Gamma(a_k)}{\Gamma(\sum_k a_k)},$$

where $\Gamma(\cdot)$ is the Gamma function.

Remark 12.1.14.

- Dirichlet distribution can be viewed as multivariate generalization of Beta distribution.
- Dirichlet distribution is usually **used as the conjugate prior** for multinomial distribution.

Lemma 12.1.30 (basic properties of Dirichlet Distribution). *Let $X = (X_1, X_2, ..., X_K), x_i \in (0, 1), \sum_{i=1}^K x_i = 1$, be a random vector with distribution $B(a), a \in \mathbb{R}^K$. Let $a_0 = \sum_{i=1}^K a_i$.*

- $$E[X_i] = \frac{a_i}{\sum_{i=1}^K a_i}.$$

- $$E[X_i^2] = \frac{a_i(a_i+1)}{(a_0)(a_0+1)}.$$
- $$Var[X_i] = \frac{a_i(a_0-a_i)}{a_0^2(a_0+1)}.$$
- *The mode of X, i.e., the value x that has the maximum probability is*

$$x_i^* = \frac{a_i - 1}{a_0 - K}.$$

Proof. (1)This can be proved using properties of Gamma distribution.

$$\begin{aligned} E[X_i] &= \int_0^1 x_i f(x) dx \\ &= \frac{\prod_{k=1}^K \Gamma(a_k + \delta_{ik})}{\Gamma(a_0+1)} / \frac{\prod_{k=1}^K \Gamma(a_k)}{\Gamma(a_0)} \\ &= \frac{a_i}{a_0} \end{aligned}$$

(2)

$$\begin{aligned} E[X_i^2] &= \int_0^1 x_i^2 f(x) dx \\ &= \frac{\prod_{k=1}^K \Gamma(a_k + 2\delta_{ik})}{\Gamma(a_0+2)} / \frac{\prod_{k=1}^K \Gamma(a_k)}{\Gamma(a_0)} \\ &= \frac{a_i(a_i+1)}{(a_0+1)a_0} \end{aligned}$$

(3) Use $Var[X] = E[X^2] - E[X]^2$. (4) To find the maximizer for $f(x)$, we take the log and maximize it. The optimality condition requires that $x_i^* \propto a_i - 1$ and $\sum_{i=1}^K a_i = 1$. □

12.1.17 χ^2-distribution

12.1.17.1 *Basic properties*

Definition 12.1.18. *A random variable X is said to have a $\chi^2(n)$ distribution with parameter $n \in \mathbb{Z}_+$ if it has pdf given as*

$$f(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-y/2}$$

with $x \in (0, +\infty)$

Remark 12.1.15 (special case of Gamma distribution). $\chi^2(n)$ has the same distribution of Gamma(n/2,2).

Definition 12.1.19 (alternative). *The χ^2-distribution with k degrees of freedom is the distribution of a sum of squares of k independent standard normal random variables. Mathematically, if $X_1, X_2, ..., X_k$ are iid random variable with $X_i \sim N(0,1)$, the random variable*

$$Q = \sum_{i=1}^{k} X_i^2$$

is distributed according to the χ^2 distribution with k degrees of freedom, writen as $Q \sim \chi^2(k)$.

Lemma 12.1.31 (basic property). *[3, pp. 161–163] Let X_1, X_2 be independent random variables. Suppose $X_1 \sim \chi^2(a_1), X_2 \sim \chi^2(a_2)$. Then*

- $Y = X_1 + X_2 \sim \chi^2(a_1 + a_2)$
- $\lambda X_1 \sim \lambda^2 \chi^2(a_1)$
- *The moment generating function is given by*

$$M(t) = (1 - 2t)^{-r/2}.$$

Proof. (1)This can be proved using properties of Gamma distribution. (2) λX_1 can be viewed as the sum of squares of normal random variables Y_i with $N(0, \lambda^2)$. Then $\sum_{i=1}^{n} (Y_i/\lambda)^2 \sim \chi^2(n)$.

□

Lemma 12.1.32 (expectation and variance). *Let random variable X has distribution of $\chi^2(n)$, then*

$$E[X] = n, Var[X] = 2n$$

In particular,

$$E[X/n] = 1, Var[X/n] = 0 \text{ as } n \to \infty.$$

that is the random variable X/n becomes deterministic constant as $n \to \infty$.

Proof. (1)Let $Z \sim \chi^2(1), Z = Y^2, Y \sim N(0,1)$, then $E[Z] = Var[Y] + (E[Y])^2 = 1$. $Var[Z] = E[Z^2] - (E[Z])^2 = E[Y^4] - 1 = 3 - 1 = 2$. (2) Use linearity of expectation that $E[X/n] = E[X]/n = 1$. Use $Var[X/n] = Var[X]/n^2 = 2/n$. □

12.1.17.2 *Noncentral chi-squared distribution*

Definition 12.1.20 (noncentral chi-squared distribution). *Let $(X_1, X_2, ..., X_k)$ be k independent, normally distributed random variables with mean μ_i and unit variances. Then the random variable*

$$Y = \sum_{i=1}^{k} X_i^2$$

*is distributed according to the **noncentral chi-squared distribution** with parameter k specifying the degree of freedom and λ, known as the **noncentrality parameter**, given by*

$$\lambda = \sum_{i=1}^{k} \mu_i^2.$$

12.1.18 Wishart distribution

Definition 12.1.21 (Wishart distribution). *Let $X_1, ..., X_n$ be independent p dimensional multivariate normal random vector with distribution $MN(0, V)$. Let $X = [X_1, ..., X_n]$. Then $M = XX^T$ is said to have Wishart distribution with parameter (n, p, V).*

Definition 12.1.22 (Wishart distribution). *A random matrix $M \in \mathbb{R}^{p\times p}$ is said to have the Wishart distribution with parameters $W_p(n, V)$ if it has pdf*

$$f(M) = \frac{1}{2^{np/2}\Gamma_p(\frac{n}{2}|V|^{n/2})}|M|^{n-p-1/2}\exp(\frac{1}{2}Tr[V^{-1}M]),$$

with the support M be the set of all symmetric positive definite matrices. Here $\Gamma_p(\alpha)$ is the multivariate gamma function.

Lemma 12.1.33 (basic properties).

- *(reduction to χ^2) If $M \in \mathbb{R}^{1\times 1}$, then*

$$M \sim W_1(n, \sigma^2) = \sigma^2\chi^2(n).$$

- *For $M \sim W_p(n, V)$, then $B^TMB \sim W_m(n, B^TVB)$, where $B \in \mathbb{R}^{p\times m}$.*
- *For $M \sim W_p(n, V)$, then $V^{-1/2}MV^{-1/2} \sim W_m(n, I)$.*
- *If M_i are independent $W_p(n_i, V)$, then $\sum_{i=1}^k \sim W_p(\sum_{i=1}^k n_i, V)$.*

- *If $M \sim W_p(n, V)$, then $E[M] = nV$.*
- *If M_1, M_2 are independent and $M_1 + M_2 = M \sim W_p(n, V)$. Further if $M_1 \sim W_p(n_1, V)$, then $M_2 \sim W_p(n - n_1, V)$.*

Lemma 12.1.34 (sample covariance). *The sample covariance*

$$\hat{Cov} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(X_i - \overline{X})^T$$

where X_i are iid $MN(0, V)$, and $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, has the property of

$$E[\hat{Cov}] = V.$$

12.1.19 t-distribution

12.1.19.1 *Standard t distribution*

Definition 12.1.23 (t distribution). *[3, p. 192] A random variable X is said to have a $t(n)$ distribution with parameter $n \in \mathbb{Z}_+$ if it has pdf given as*

$$f(x) = \frac{\Gamma((n+1)/2)}{\Gamma(n/2)\sqrt{\pi n}}(1 + \frac{t^2}{n})^{-(n+1)/2}$$

with $x \in (-\infty, +\infty)$

Definition 12.1.24 (alternative). *Let random variable $W \sim N(0,1)$, Let random variable $V \sim \chi^2(n)$ **independent of** V. Define a new random variable T as*

$$T = \frac{W}{\sqrt{V/n}}$$

Then T has a t-distribution with degree of freedom n, denoted by T_n or t_n.

Remark 12.1.16 (comparison with normal distribution).

- t distribution generally have shorter peak and fatter tails than normal distribution.
- $t_n \to N(0,1)$ as $n \to \infty$.

Lemma 12.1.35 (mean and variance of t-distribution). *The mean for a t-distribution with degree of n is given by*

$$E[t_n] = \begin{cases} 0, n > 1 \\ \infty(undefined), n = 1 \end{cases} .$$

The variance for a t-distribution with degree of n is given by

$$Var[t_n] = \begin{cases} \frac{n}{n-2}, n > 2 \\ \infty, n = 1, 2 \end{cases} .$$

12.1.19.2 *classical t distribution*

Definition 12.1.25. *[8, p. 95] If Y has a standard t_n distribution, then*

$$Z = \mu + \lambda Y$$

is said to have a $t_n(\mu, \lambda^2)$ distribution.

Lemma 12.1.36 (mean and variance of classical t-distribution). *Let Z be a random variable $t_n(\mu, \lambda^2)$. Then*

$$E[Z] = \begin{cases} \mu, n > 1 \\ \infty(undefined), n = 1 \end{cases} ,$$

and

$$Var[Z] = \begin{cases} \lambda^2 \frac{n}{n-2}, n > 2 \\ \infty, n = 1, 2 \end{cases} .$$

12.1.19.3 *Multivariate t distribution*

Definition 12.1.26 (multivariate t distribution). *[8]*

- *Let Z be a d dimensional multivariate Gaussian $MN(0,\Sigma)$, and $\mu \in \mathbb{R}^d$. The d dimensional random vector Y, defined as,*

$$X = \mu + \sqrt{\frac{n}{W}}Z,$$

*where $W \sim \chi^2(n)$ and W is **independent** of Z, has a $t_n(\mu,\Sigma)$ multivariate distribution.*
- *Let $X \sim t_n(\mu,\Sigma)$. Then X has the density given by*

$$f(x) = \frac{\Gamma((n+d)/2)}{\Gamma(n/2)n^{d/2}\pi^{d/2}|\Sigma|^{1/2}}(1+\frac{1}{n}(x-\mu)^T\Sigma^{-1}(x-\mu))^{-(n+d)/2}.$$

Lemma 12.1.37 (mean and variance of multivariate t-distribution). *Let Z be a random variable $t_n(\mu,\Sigma)$. Then*

$$E[Z] = \begin{cases} \mu, n > 1 \\ \infty(undefined), n = 1 \end{cases},$$

and

$$Cov[Z] = \begin{cases} \Sigma\frac{n}{n-2}, n > 2 \\ \infty, n = 1,2 \end{cases}.$$

12.1.20 F-distribution

Definition 12.1.27 (F distribution). *[3, p. 192] A random variable X is said to have a $F(n_1,n_2)$ distribution with parameter $n_1,n_2 \in \mathbb{Z}_+$ if it has pdf given as*

$$f(x) = \frac{\Gamma((n_1+n)2)/2)(n_1/n_2)^{n_1/2}y^{n_1/2-1}}{\Gamma(n_1/2)\Gamma(n_2/2)[1+(n_1x/n_2)]^{(n_1+n_2)/2}}$$

with $x \in (0,+\infty)$

Definition 12.1.28 (alternative). *Given two **independent** chi-squared random variables W and V having r_1 and r_2 degrees of freedom. We define a new random variable*

$$W = \frac{U/r_1}{V/r_2}$$

Then W has a F-distribution with parameter (r_1, r_2).

Lemma 12.1.38 (inverse relationship). *Let X be a random variable with distribution $F(n_1, n_2)$, then $1/X$ is a random variable with distribution $F(n_2, n_1)$.*

Proof. Directly from definition. □

Lemma 12.1.39 (relationship to t distribution). *Let X be a random variable with standard t distribution with n degrees of freedom. Then*

$$X^2 \sim F(1, n).$$

That is, X^2 has the distribution of $F(1, n)$.

Proof. Directly from definition. □

Definition 12.1.29 (noncentral F distribution). *Given two chi-squared random variables W and V such that V is a noncentral chi-squared random variable with non-centrality parameter λ and degree of freedom r_1, and W is a chi-squared random variable having r_1 r_2 degrees of freedom. We define a new random variable*

$$W = \frac{U/r_1}{V/r_2}$$

Then W has a noncentral F-distribution with parameter (λ, r_1, r_2). .

12.1.21 Empirical distributions

Definition 12.1.30 (empirical cumulative distribution function(CDF)). *Given N iid random variables $Y_1, Y_2, ..., Y_N$ with common cdf $F(t)$, the empirical CDF is defined by*

$$\hat{F}_N(t) = \frac{number\ of\ elements\ in\ the\ sample \leq t}{N} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{Y_i \leq t}$$

.

Lemma 12.1.40 (basic statistic properties). *Let $\hat{F}_N(t)$ be the empirical cdf of a random sample of size N. For a fixed t, we have*

- *$N\hat{F}_N(t)$ is a binomial random variable with parameter (N, p),where $p = F(t)$.*
- *$N\hat{F}_N(t)$ is an unbiased estimator for $NF(t)$.*
- *$N\hat{F}_N(t)$ has variance $NF(t)(1 - F(t))$.*

Proof. (1) Note that based on the definition of $\hat{F}_N(t) \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{Y_i \leq t}$, $\mathbf{1}_{Y_i \leq t}$ is a Bernoulli random variable with parameter $p = F(t)$. Therefore, $N\hat{F}_N(t) = \sum_{i=1}^{N} \mathbf{1}_{Y_i \leq t}$ will follow a binomial distribution of parameter (N, p). (2)

$$E[N\hat{F}_N(t)] = Np = NF(t).$$

(3)

$$Var[N\hat{F}_N(t)] = Np(1 - p) = NF(t)(1 - F(t)).$$

□

12.1.22 Heavy-tailed distributions

12.1.22.1 *Basic characterization*

Definition 12.1.31 (Heavy-tailed distribution). *The distribution of a random variable X with distribution function F is said to have a heavy right tail if*

$$\lim_{x \to \infty} e^{\lambda x} Pr(X > x) = \infty, \forall \lambda > 0.$$

Remark 12.1.17 (interpretation). Heavy-tailed distributions have densities decaying slower in the tails than the normal.

12.1.22.2 *Pareto and power distribution*

Definition 12.1.32 (Pareto distribution). *A random variable X is said to have Pareto distribution with scale parameter $x_m > 0$ and shape parameter $\alpha > 0$ if its has pdf*

$$f_X(x) = \begin{cases} \frac{\alpha x_m^{\alpha}}{x^{\alpha+1}}, x \geq x_m, \\ 0, x < x_m. \end{cases};$$

or cdf

$$f_X(x) = \begin{cases} 1 - (\frac{\alpha x_m}{x})^\alpha, x \geq x_m, \\ 0, x < x_m. \end{cases}.$$

X has support $[x_m, \infty)$.

Definition 12.1.33 (power law distribution). *A random variable X is said to have power law distribution with parameters* K, α *if its has probability characterization on its tail given by*

$$Pr(X > x) = Kx^{-\alpha}.$$

Remark 12.1.18 (Pareto distribution and power law distribution are heavy-tailed distribution). Note that since power grows much slower than the exponentialA.2.1, therefore

$$\lim_{x \to \infty} e^{\lambda x} Pr(X > x) = \infty, \forall \lambda > 0.$$

12.1.22.3 *Student t distribution family*

Definition 12.1.34 (Student's t-Distribution family). *The t distribution has a single parameter,* $\nu > 0$, *known as* degrees of freedom. *The density function is given as*

$$f_\nu(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}(1 + \frac{x^2}{\nu})^{-\frac{1}{2}(\nu+1)}$$

The first two members of family are

1. $f_1(x) = \frac{1}{\pi(1+x^2)}$
2. $f_2(x) = \frac{1}{2\sqrt{2}}(1 + x^2/2)^{-3/2}$

The $\nu = 1$ *density is known as* Cauchy's density. *As* $\nu \to \infty$, *the density distribution tends to the standard normal density.*

Definition 12.1.35 (Cauchy distribution). *The Cauchy distribution with parameter* (x_0, γ) *has the probability density function*

$$f(x; x_0, \gamma) = \frac{1}{\pi\gamma}(\frac{\gamma^2}{(x - x_0)^2 + \gamma^2}),$$

where x_0 is the location parameter, specifying the location of the peak of the distribution, and γ is the scale parameter which specifies the half-width at half-maximum. ***Standard Cauchy distribution*** *is Cauchy distribution with parameter* $(0, 1)$.

Remark 12.1.19 (nonexistence of moments).

- The Cauchy distribution is an example of a distribution which has no mean, variance or higher moments. And therefore the moment generating function does not exist. However, the **mode and median** are well defined and both equal to x_0.
- The nonexistence of expectation is because of the $E[|X|] < \infty$.

Lemma 12.1.41 (sum of Cauchy distribution). *If $X_1, ..., X_n$ are independent and identically distributed random variables, each with a standard Cauchy distribution, then the sample mean*

$$\overline{X} = (X_1 + ... + X_n)/n$$

has the same standard Cauchy distribution.

Proof. Note that we need to use characteristic function to prove, since the moment generating function does not exist. □

12.1.22.4 *Gaussian mixture distributions*

Definition 12.1.36 (normal scale mixture distribution). *[8, p. 99] The normal scale mixture distribution is the distribution of the random variable*

$$Y = \mu + \sqrt{U}Z,$$

where μ is constant equal to the mean, and $Z \sim N(0, 1)$, U is a positive random variable giving the variance of each component, and Z and U are independent.

If U can assume only a finite number of values, then Y has a ***discrete scale mixture distribution****. If U is continuously distributed, then Y has a* ***continuous scale mixture distribution****.*

Example 12.1.7 (discrete Gaussian mixture distribution). Let $\mu = 0$, and U have the following distribution

$$P(U = 25) = 0.1, P(U = 1) = 0.9.$$

Then

$$Y = \mu + \sqrt{U}Z,$$

is the mixture of 10% of $N(0, 25)$ and 90% of $N(0, 1)$.

Example 12.1.8 (t distribution). The t_n distribution with n degrees of freedom is a continuous Gaussian mixture with

$$\mu = 0, U = \frac{n}{W},$$

where $W \sim \chi^2(n)$.

Definition 12.1.37 (multivariate normal variance mixtures). *The random vector X has a multivariate normal variance mixture distribution if*

$$X \triangleq \mu + \sqrt{W}AZ$$

where

- $Z \sim MN(0, I_k)$
- *W is a **positive** scalar random variable which is independent of Z*
- *$A \in \mathbb{R}^{d \times k}$ and $\mu \in \mathbb{R}^d$ are a matrix and a vector of constants*

Example 12.1.9 (special case: multivariate t distribution). The t_n distribution with n degrees of freedom is a continuous Gaussian mixture with

$$\mu = 0, U = \frac{n}{W},$$

where $W \sim \chi^2(n)$.

12.2 Characterizing distributions

12.2.1 Skewness and kurtosis

Skewness is a measure of symmetry of a statistical distribution [Figure 12.2.1].There are two types of skewness.

- **Negative skewness** indicates that the mean of the data values is less than the median, and the data distribution is **left-skewed**.
- **Positive skewness** indicates that the mean of the data values is greater than the median, and the data distribution is **right-skewed**.

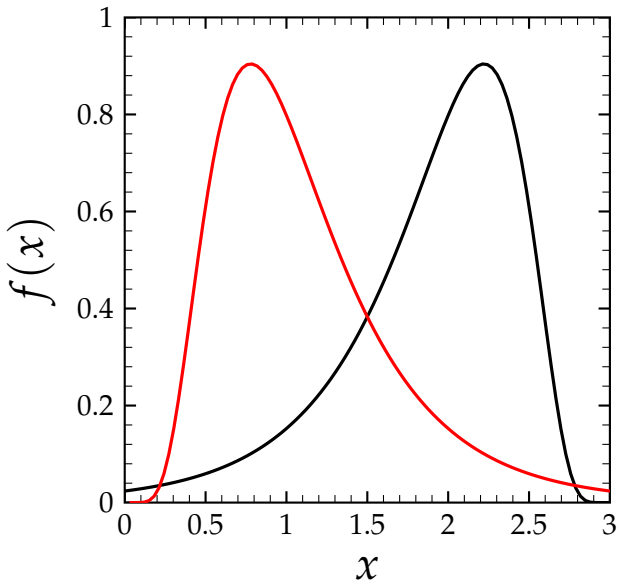


Figure 12.2.1: Distributions with left-skewness (black) and right-skewness (red).

Skewness can be computed quantitatively via the following definition.

Definition 12.2.1 (skewness). *The skewness of an univariate population for random variable X is defined by*

$$\gamma_1 = E[(\frac{X-\mu}{\sigma})^3] = \frac{E[(X-\mu)^3]}{(E[(X-\mu)^2])^{3/2}} = \frac{\mu_3}{\mu_2^{3/2}}$$

*where μ_2 and μ_3 are the second and the third **central moments**.*

Example 12.2.1. Let $X \sim N(\mu, \sigma^2)$. Then the skewness of X distribution is $\gamma_1 = E[(\frac{X-\mu}{\sigma})^3] = E[Z^3] = 0, Z \sim N(0,1)$, where we use the fact the third moment for a standard normal is zero [Lemma 12.1.10].

Kurtosis is a measure of tail shape of a distribution. There are three types of kurtosis [Figure 12.2.2]:

- **Mesokurtic** distributions have zero excess kurtosis. Normal distribution is mesokurtic.
- **Leptokurtic** distributions have excess kurtosis greater than 0. This type of distribution is one with extremely thick tails and a very thin and tall peak. t-distribution and Laplace distribution are leptokurtic.
- **Platykurtic** distributions have excess kurtosis smaller than 0. This type of distribution has a short and broad-looking peak. Uniform distribution is platykurtic.

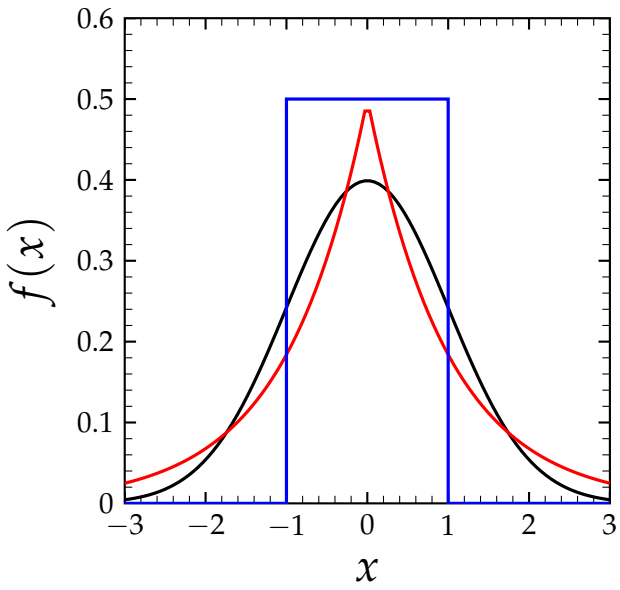


Figure 12.2.2: Distributions with zero excess Kurtosis (Normal distribution, black), positive excess kurtosis (Laplace distribution, red), and negative excess Kurtosis (Uniform distribution, blue).

Kurtosis can be computed quantitatively via the following definition.

Definition 12.2.2 (kurtosis, excess kurtosis).

- *The **kurtosis** of a univariate population is defined by*

$$\gamma_2 = E[(\frac{X-\mu}{\sigma})^4] = \frac{E[(X-\mu)^4]}{(E[(X-\mu)^2])^2} = \frac{\mu_4}{\mu_2^2},$$

  where μ_2 and μ_4 are the second and the fourth central moments.
- *The **excess kurtosis** of a univariate population is defined by*

$$\gamma_2^{ex} = \gamma_2 - 3.$$

Example 12.2.2. Let $X \sim N(\mu, \sigma^2)$. Then the Kurtosis of X distribution is $\gamma_2 = E[(\frac{X-\mu}{\sigma})^4] = E[Z^4] = 3, Z \sim N(0,1)$, where we use the fact the fourth moment for a standard normal is 3 [Lemma 12.1.10].

12.2.2 Percentiles and quantiles

12.2.2.1 *Basics*

Percentiles are cut points in the support that divide the distribution into different parts, with each part occupying different probability mass. Compared to skewness and kurtosis, percentiles offers a more comprehensive characterization on the shape of the distribution.

Definition 12.2.3 (percentile of a distribution). *The α **percentile**($\alpha \in [0,1]$) of a probability distribution of random number X is a number p in the support D of the support such that*

$$Pr(x < p) = \alpha, Pr(x > p) = 1 - \alpha.$$

Or equivalently, the α percentile is given by the inverse of cdf

$$p = F_X^{-1}(\alpha).$$

Example 12.2.3 (percentiles of a standard normal distribution). In Figure 12.2.3, we plot percentiles at $\alpha = 0.1, 0.2, ..., 0.9$ for a standard normal distribution.

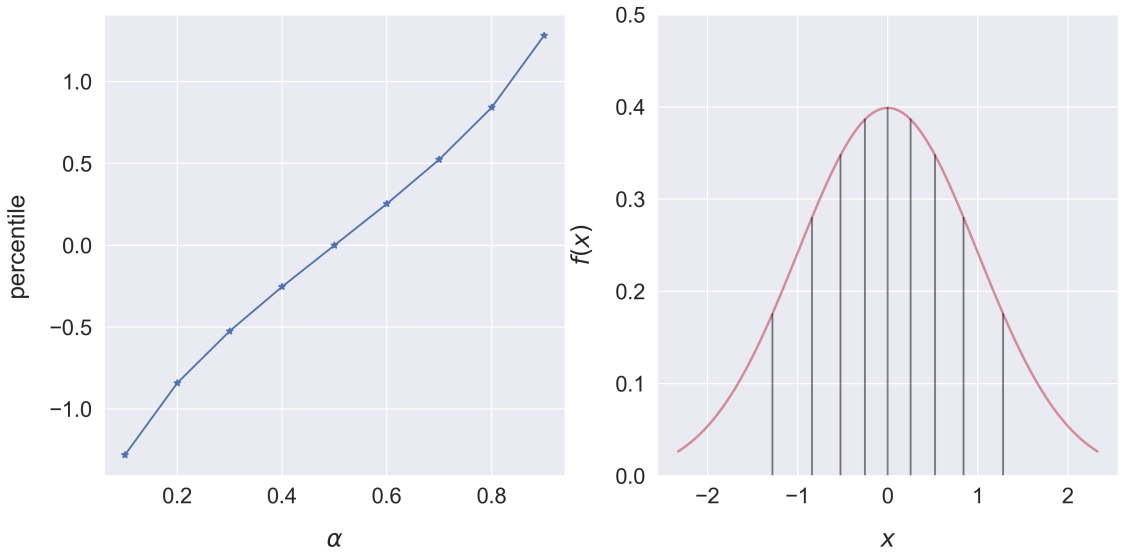


Figure 12.2.3: Percentile points at $\alpha = 0.1, 0.2, ..., 0.9$ for a standard normal distribution.

Definition 12.2.4 (percentile in a set of sample values). *The α **percentile** ($\alpha \in [0,1]$) of a set of values is a value in $\mathbb{R}$ that divides them so that $100\alpha\%$ of values lie below and $100(1-\alpha)\%$ of the values lie above.*

The calculation of a α percentile in N sample values is quite straight forward:

- Sort the N values in ascending order.
- The number located at $N\alpha$ (rounded integer) is the percentile.

Quantiles are the cut points dividing the range of a probability distribution into contiguous intervals with equal probabilities. We can visually compare if two distributions are similar by plotting the quantiles of two distributions against each other, known as **QQ plot**. In particular, if one random variable is the affine transformation of the other random variable, then the percentiles of the former is the affine transformation of the latter, as showed in the following.

Lemma 12.2.1 (linear relationship between percentiles from two distributions). *Let X and Y be two random variables with cdf F_X and F_Y. Let $p_X = F_X^{-1}(\alpha)$ and $p_Y = F_Y^{-1}(\alpha)$ for $\alpha \in [0,1]$. It follows that*

- *If $Y = aX + b$, then*
$$p_Y = ap_X + b$$
- *If $Y = \alpha X^\beta$, then*
$$p_{\ln Y} = \beta p_{\ln X} + \ln \alpha,$$

where $p_{\ln Y} = F^{-1}_{\ln Y}(\alpha), p_{\ln X} = F^{-1}_{\ln X}(\alpha)$,

Proof. (1) We know that

$$\alpha = F_Y(p_Y) = F_X(p_X).$$

From scale-location transformation [Lemma 11.4.7], we have

$$p_X = (p_Y - b)/a.$$

(2) From $Y = \alpha X^\beta$, we have $\ln Y = \beta \ln X + \ln \alpha$. □

In Figure 12.2.4, we demonstrate the usage of QQ plot to compare different sample distributions as standard normal distribution. As expected, for samples drawn from a normal distribution or a shifted-scaled normal distribution, the QQ plot is approximately a perfect straight line; For Student's t distribution, which has a heavier tails than normal distribution, we see the deviation from the straight line at the two ends; For a log-normal distribution, which is vastly different from the normal distribution, we see strong deviations accordingly.

12.2.2.2 *Cornish-Fisher expansion*

With the knowledge of skewness and kurtosis, we can also approximate quantiles of a random variable using following **Cornish-Fisher expansion**.

Theorem 12.2.1 (Cornish-Fisher expansion). *Consider a distribution with mean μ and variance σ^2. Then its α quantile can be approximate by*

$$\mu + \sigma z_\alpha^{cf}$$

where

$$z_\alpha^{cf} = q_\alpha + \frac{(q_\alpha^2 - 1)S(X)}{6} + \frac{(q_\alpha^3 - 3q_\alpha)K(X)}{24} - \frac{(2q_\alpha^3 - 5q_\alpha)S^2(X)}{36},$$

where $S(X)$ is skewness, $K(X)$ is kurtosis, z_α^{cf} is the Cornish-Fisher approximate quantile value for the confidence level α, and q_α is the quantile value for the standard normal distribution with confidence level α.

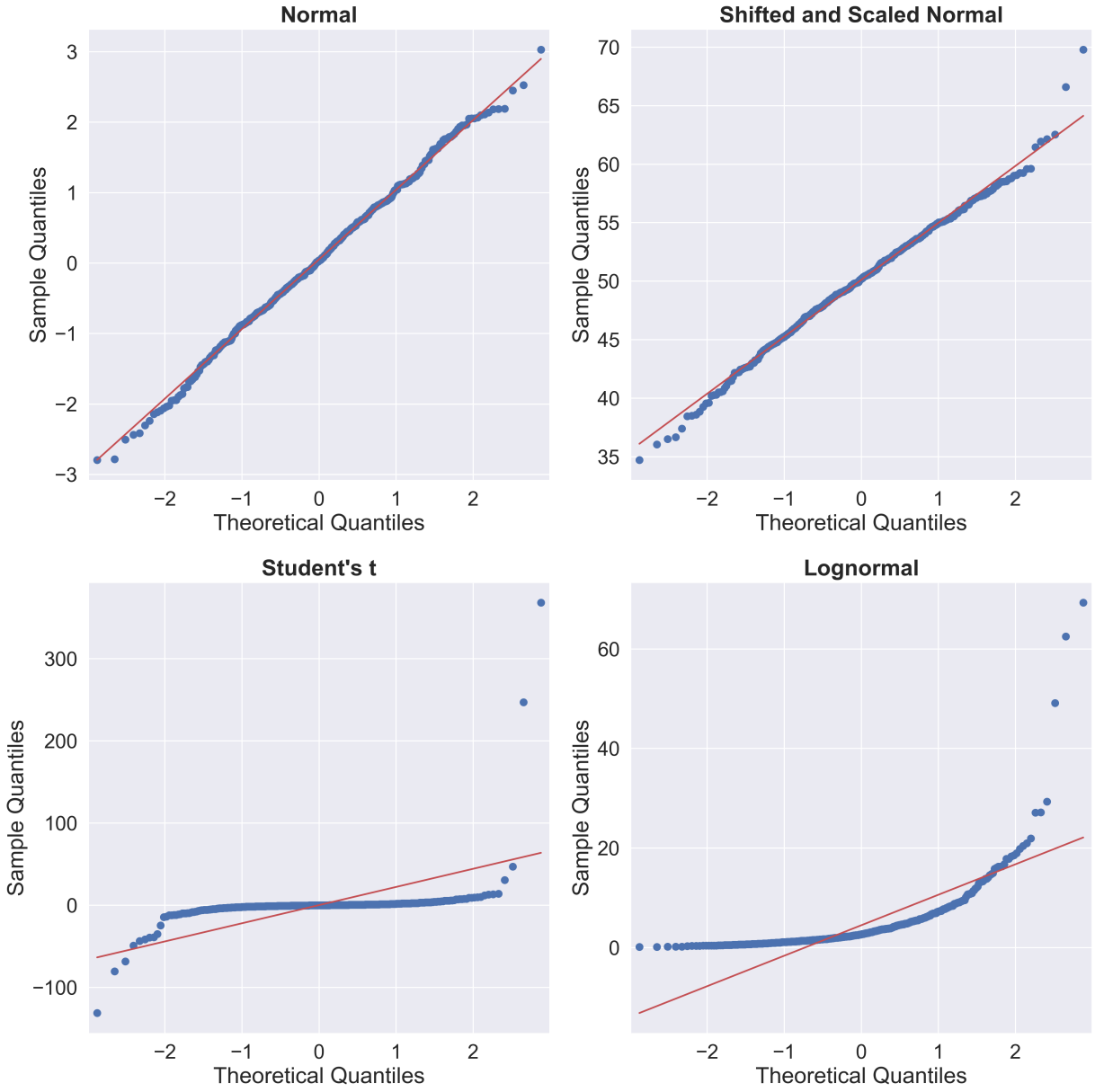


Figure 12.2.4: QQ plot of different sample distributions against standard normal distribution, including standard normal $N(0,1)$, shifted-scaled normal $N(50,5)$, Student's t with degree 1, and lognormal $LN(0,1)$. Red solid lines are the fitted linear lines.

12.3 Moment matching approximation methods

In our previous sections, we have covered the mostly widely used statistical distributions. In real-world applications, the statistical distribution of a random quantity is in general unknown. A practical modeling approach is to approximate the distribution by these discrete or continuous parametric distributions. Arguably the most common used, and perhaps simplest, approximate distribution is normal distribution, which is largely justified by the Central Limit Theorem [Theorem 11.11.3].

However, normal distributions also have their limitations: first, normal distributions do not have a bounded support, which is not ideal in modeling distribution with bounds (e.g., range greater than 0); second, normal distributions have zero skewness, making it a bad modeling choice for skewed distributions.

In this section, we consider the log-normal distribution family as an alternative to normal distribution. In subsection 12.1.10, we see that log-normal distributions have different extensions and mulitple parameters to control the boundedness and skewness; further, log-normal distributions have excellent analytical tractability.

Using log-normal distribution as an example, here we discuss the moment matching method to determine distribution parameters.

Lemma 12.3.1 (2 parameter log-normal approximation via moment matching). *Suppose we have a random variable X with range $X > 0$. Suppose X has moments given by*

$$E[X] = M_1, E[X^2] = M_2.$$

Let Y be a log-normal random variable defined by

$$Y = M_1 \exp(-\frac{1}{2}v^2 + vZ), Z \in N(0,1),$$

where

$$v^2 = \log(M_2/M_1^2)$$

Then Y has the same first two moments as X; that is

$$E[Y] = M_1, E[Y^2] = M_2.$$

Proof. Using moment generating function of Z, we know that

$$E[Y] = M_Z(v)M_1 \exp(-\frac{1}{2}v^2) = M_1.$$

and

$$E[Y^2] = M_Z(2v)M_1^2 \exp(-v^2) = \exp(v^2)M_1^2 = \frac{M_2}{M_1^2}M_1^2 = M_2.$$

□

Similarly, this moment matching method can be generalized to there-parameter log-normal distributions.

Lemma 12.3.2 (3 parameter shifted lognormal approximation via moment matching). *Suppose we have a random variable X having moments given by*

$$E[X] = M_1, E[X^2] = M_2, E[X^3] = M_3.$$

Let Y be a shifted log-normal random variable with parameter $SLN(\mu, \sigma^2, \tau)$ such that

$$E[Y] = \tau + \exp(\mu + \frac{1}{2}\sigma^2),$$

$$E[Y^2] = \tau^2 + 2\tau \exp(\mu + \frac{1}{2}\sigma^2) + \exp(2\mu + 2\sigma^2),$$

$$E[Y^3] = \tau^3 + 3\tau^2 \exp(\mu + \frac{1}{2}\sigma^2) + 3\tau \exp(2\mu + 2\sigma^2) + \exp(3\mu + \frac{9}{2}\sigma^2).$$

If we can find (μ, σ, τ) such that

$$E[X] = E[Y], E[X^2] = E[Y^2], E[X^3] = E[Y^3],$$

then X and Y have matched moments.

Proof. For moments of Y, see Lemma 12.1.20. □

Based on the target distribution's location and skewness, we can choose the type of lognormal distribution we want to use. The table below is a good summary[4].

skewness	$\gamma > 0$	$\gamma > 0$	$\gamma < 0$	$\gamma < 0$
location	$\tau \geq 0$	$\tau < 0$	$\tau \geq 0$	$\tau < 0$
choice of approximation	regular	shifted	negative	negative shifted

12.4 Gaussian quadratic forms

12.4.0.1 *Quadratic forms and chi-square distribution*

Let $X = (X_1, X_2, ..., X_n)^T$ be a random vector, we called

$$Q = X^T \Sigma X, \Sigma \in \mathbb{R}^{n \times n},$$

a **quadratic form of random vector** X. Note that Q is also a random variable. We are particularly interested in the case where X follows a multivariate Gaussian distribution. In this case, Q is known as **Gaussian quadratic form**.

Gaussian quadratic forms are widely used in characterizing residual or error distributions in linear regression applications. We start with its basic property.

Lemma 12.4.1. *Let X be a m-dimensional random vector with multivariate Gaussian distribution, i.e., $X \sim N(\mu, \Sigma)$. It follows that*

- $$\Sigma^{1/2}(x - \mu) \sim N(0, I).$$
- $$(x - \mu)\Sigma^{-1}(x - \mu) \sim \chi^2(m).$$

Proof. (1) Directly from affine transformation property of multivariate Gaussian random variable [Theorem 14.1.1]. (2) Use the definition that sum of iid normal random variable square is chi-square random variable. □

Theorem 12.4.1 (chi-square orthogonal decomposition). *Let $X_1, X_2, ..., X_n$ be independent standard normal variables such that*

$$\sum_{i=1}^{n} X_i^2 \sim \chi^2(n).$$

Denote $X = (X_1, ..., X_n)^T$. If there exists an orthogonal projector $P \in \mathbb{R}^{n \times n}$ such that $Y = PX, Z = (I - P)X$, then

- *$Y \sim MN(0, P), Z \sim MN(0, I - P)$, and Y, Z are independent of each other.*
- *$Y^T Y \sim \chi^2(r), r = rank(P)$; or equivalently, the quadratic form $Q = X^T P X \sim \chi^2(r)$.*
- *$Z^T Z \sim \chi^2(n - r)$;or equivalently, the quadratic form $Q = X^T (I - P)X \sim \chi^2(n - r)$*

In summary, for a quadratic form $Q = X^T \Sigma X$, if Σ is idempotent and symmetric, then $Q \sim \chi^2(rank(\Sigma))$.

Proof. (1) From affine transform of multivariate normal [Theorem 14.1.1],

$$Y \sim MN(0, P\Sigma_X P^T) = MN(0, P^2) = MN(0, P).$$

To show independence, we have $E[YZ^T] = E[PXX^T(I-P)^T] = E[P(I-P)] = 0$.

(2) Let U be the eigen-decomposition of P such that $P = UU^T$. Let $Z = U^T X, Z \in \mathbb{R}^r, Z \sim MN(0, I_r)$. Let V be the eigen-decomposition of $I - P$ such that $I - P = VV^T$. Let $W = V^T X, W \in \mathbb{R}^{n-r}, W \sim MN(0, I_{n-r})$. We want to show that the characteristic function of the random quantity $Y^T Y$ is the same as the characteristic function of $\chi^2(r)$.

$$\begin{aligned}
&E[\exp(itY^T Y)] \\
=&\frac{1}{(2\pi)^{n/2}} \int\int\cdots\int \exp(it(U^T X)^T (U^T X) \exp(-\frac{1}{2} X^T (I - P + P)X)) dx_1 dx_2 \cdots dx_n \\
=&\frac{1}{(2\pi)^{n/2}} \int\int\cdots\int \exp(itZ^T Z) \exp(-\frac{1}{2}(Z^T Z + W^T W)) dz_1 \cdots dz_r dw_{r+1} \cdots dw_n \\
=&\frac{1}{(2\pi)^{r/2}} \int\int\cdots\int \exp(itZ^T Z) \exp(-\frac{1}{2}(Z^T Z)) dz_1 \cdots dz_r
\end{aligned}$$

where we change the integral variable such that

$$[dz_1 \cdots dz_r dw_{r+1} \cdots dw_n]^T = [U\ V](dx_1 dx_2 \cdots dx_n)^T$$

. The last line is the characteristic function of $\chi^2(r)$. (3) similar to (2). □

Lemma 12.4.2 (moment generating functions for Gaussian quadratic forms). *[3, p. 523] Let $X = (X_1, X_2, ..., X_n)^T$ where $X_1, X_2, ..., X_n$ are iid $N(0,1)$. Consider the quadratic form $Q = X^T AX$ for a symmetric matrix A of rank $r \leq n$. It follows that*

- *Q has the moment generating function $M(t) = \prod_{i=1}^r (1 - 2t\lambda_i)^{-1/2} = |I - 2tA|^{-1/2}$, where $\lambda_1, \lambda_2, ..., \lambda_r$ are the nonzero eigenvalues of $A, |t| < \frac{1}{\max|\lambda_i|}$.*
- *If A is an orthogonal projector such that $\lambda_1 = \lambda_2 = \cdots = \lambda_r = 1$, then*

$$M(t) = M_{\chi^2(r)}.$$

Proof. (1)Let the eigen-decomposition of A be

$$A = U\Lambda U^T, U \in \mathbb{R}^{n \times r}, \Lambda \in \mathbb{R}^{r \times r}.$$

Then

$$Q = X^T AX = X^T U \Lambda U^T X = X^T (\sum_{i=1}^{r} \lambda_i u_i u_i^T) X = \sum_{i=1}^{r} \lambda_i (u_i^T X)^T.$$

Let $Y_i = u_i^T X, i = 1, 2, ..., r$. It can be shown that $Y_i \sim N(0,1), E[Y_i Y_j] = u_i^T E[XX^T] u_j^T = \delta_{ij}$; that is $Y_1, Y_2, ..., Y_r \sim MN(0, I_r)$. Therefore, $Y_i^2 \sim \chi^2(1)$.

The moment generating function is given by

$$\begin{aligned} M(t) &= E[\exp(tQ)] \\ &= E[\exp(t \sum_{i=1}^{r} \lambda_i Y_i^2)] \\ &= \prod_{i=1}^{r} E[\exp(t \lambda_i Y_i^2)] \\ &= \prod_{i=1}^{r} M_{\chi^2(1)}(\lambda_i t) \\ &= \prod_{i=1}^{r} (1 - 2\lambda_i t)^{-1/2} \end{aligned}$$

where we use the moment generating function of $\chi^2(1)$ from Lemma 12.1.31. (2) straight forward. □

Lemma 12.4.3 (independence of quadratic forms). *[3, p. 528] Let $X = (X_1, X_2, ..., X_n)$ be a random vector where $X_1, X_2, ..., X_n$ are iid $N(0,1)$. For real symmetric matrices $A, B \in \mathbb{R}^{n \times n}$, let $Q_1 = X^T AX$ and $Q_2 = X^T BX$. Then Q_1 and Q_2 are independent if and only if $AB = 0$.*

Proof. Let $rank(A) = r, rank(B) = s$. Let the eigendecomposition of A, B be such that

$$A = \sum_{i=1}^{r} \lambda_i u_i u_i^T, A = \sum_{i=1}^{s} \beta_i v_i v_i^T.$$

If $AB = 0$, then $u_1, ..., u_r, v_1, ..., v_r$ will be orthogonal to each other. Then

$$Q_1 + Q_2 = \sum_{i=1}^{r+s} \lambda_i u_i u_i^T,$$

where $u_{r+i} = v_i, \lambda_{r+i} = \beta_i$.

It is easy to see that [Lemma 12.4.2]

$$M_{Q_1, Q_2}(t_1, t_2) = M_{Q_1}(t_1) M_{Q_2}(t_2).$$

Then from independence-from-mgf [Lemma 11.6.3], we can prove Q_1 and Q_2 are independent. □

12.4.0.2 *Applications*

The first application is to prove Student's theorem, which specifies the distribution of sample variance.

Theorem 12.4.2 (Student's Theorem). *Let $X_1, X_2, ..., X_n$ be iid random variables each having a normal distribution with mean μ and variance σ^2. Define random variables as:*

$$\bar{X} = \frac{1}{n}\sum_{i=0}^{n} X_i, S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

1. *$\bar{X}$ has a $N(\mu, \sigma^2/n)$ distribution*
2. *$\bar{X}$ and S^2 are independent.*
3. *$(n-1)S^2/\sigma^2$ has a $\chi^2(n-1)$ distribution*
4. *The random variable*

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

has t-distribution with $n-1$ degrees of freedom.

Proof. (1)From Lemma 12.1.9. (2) We can prove $\overline{X}$ and the random vector $Y = (X_1 - \overline{X}, ..., X_n - \overline{X})$ are independent. Note that

$$\overline{X} = \frac{1}{n}\mathbf{1}^T X, Y = (I - \frac{1}{n}\mathbf{1}\mathbf{1}^T)X,$$

and hence $\overline{X}$ and Y are both normal.

$$\begin{aligned} Cov(\overline{X}, Y) &= X^T(\frac{1}{n}\mathbf{1}^T(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T))X \\ &= X^T\frac{1}{n}(\mathbf{1}^T - \frac{1}{n}\mathbf{1}^T\mathbf{1}\mathbf{1}^T)X \\ &= X^T\frac{1}{n}(\mathbf{1}^T - \mathbf{1}^T))X = 0 \end{aligned}$$

where we use the fact that $\mathbf{1}^T\mathbf{1} = n$.

Then $S^2 = \frac{1}{n-1}Y^TY$ will be independent of $\overline{X}$ because S^2 is a function of Y [Lemma 11.3.2]. (3) See reference and Corollary 12.4.3.1. (4) From the definition of the t distribution, we have

$$Y = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

is the $N(0,1)$. $W = (n-1)S^2/\sigma^2$ has a $\chi^2(n-1)$ distribution. Then

$$\frac{Y}{\sqrt{W/n-1}}$$

has $t(n-1)$ distribution. □

The second application is to prove Cochran's theorem.

Lemma 12.4.4. *Let $X_1, X_2, ..., X_n$ be real numbers. Suppose that $\sum_{i=1}^n X_i^2$ can be decomposed into a sum of positive semi-definite quadratic forms, that is*

$$\sum_{i=1}^n X_i^2 = Q_1 + ... + Q_k$$

where $Q_i = X^T A_i X$ with $rank(A_i) = r_i$. If $\sum_{i=1}^k r_i = n$, then there exists an orthonormal matrix C such that $X = CY$ and

$$\begin{aligned} Q_1 &= Y_1^2 + ... + Y_{r_1}^2 \\ Q_2 &= Y_{r_1+1}^2 + ... + Y_{r_1+r_2}^2 \\ &\cdots \end{aligned}$$

Proof. (informal)Note that when we decompose a matrix, its sum of rank of the decomposed matrix will increase [Theorem 4.4.1], i.e.,

$$rank(A+B) \leq rank(A) + rank(B)$$

and the equality only holds when $\mathcal{R}(A) \cap \mathcal{R}(B) = \emptyset$.
Since in our case $rank(\sum A_i) = \sum rank(A_i)$, then we must have $\mathbb{R}^n = \mathcal{R}(A_1) \oplus \mathcal{R}(A_2) ... \oplus \mathcal{R}(A_k)$. Take the basis of each $\mathcal{R}(A_i)$ and make it to be orthonormal matrix C. Then Y_i are just the orthonormal projection to subspace $\mathcal{R}(A_j)$. An accessible proof is at Theorem 12.4.1 □

Theorem 12.4.3 (Cochran's theorem). *Let $X_1, X_2, ..., X_n$ be iid $N(0, \sigma^2)$ random variables. Suppose that $\sum_{i=1}^n X_i^2$ can be decomposed into a sum of positive semi-definite quadratic forms, that is*

$$\sum_{i=1}^n X_i^2 = Q_1 + ... + Q_k$$

where $Q_i = X^T A_i X$ with $rank(A_i) = r_i$. If $\sum_{i=1}^k r_i = n$, then there exists an orthonormal matrix C such that $X = CY, Y = C^T X$ ($Y_1, Y_2, ..., Y_n$ are independent random variables with $N(0, \sigma^2)$)and

$$\begin{aligned} Q_1 &= Y_1^2 + ... + Y_{r_1}^2 \\ Q_2 &= Y_{r_1+1}^2 + ... + Y_{r_1+r_2}^2 \\ &\dots \end{aligned}$$

Moreover, we have

- *$Q_1, Q_2, ..., Q_k$ are independent*
- *$Q_i \sim \sigma^2 \chi^2(r_i)$.*

Proof. (1) Use above lemma. Note that $Y_1, Y_2, ..., Y_n$ are still independent normal because of Lemma 14.1.2. (2) Since Q_i and Q_j have non-overlapping Y_is, they are independent to each other. (3) From properties of χ^2 distribution [Lemma 12.1.31]. □

Corollary 12.4.3.1 (distribution of sample variance). *Let $Y_1, ..., Y_n$ be iid random variable with $N(\mu, \sigma^2)$, then*

$$\sum_{i=1}^n (Y_i - \overline{Y})^2 \sim \sigma^2 \chi^2(n-1)$$

and

$$\sum_{i=1}^n (Y_i - \mu)^2 / n \sim \sigma^2 \chi^2(1)$$

Proof.

$$\sum_{i=1}^n (Y_i - \mu)^2 = \sum_{i=1}^n (Y_i - \overline{Y})^2 + \sum_{i=1}^n (Y_i - \mu)^2 / n$$

And

$$(Y-\mu)^T (Y-\mu) = (Y-\mu)^T (I - \frac{1}{n} J)(Y-\mu) + (Y^T - \mu)^T (\frac{1}{n} J)(Y-\mu)$$

and $rank(\frac{1}{n}J)$ has rank 1 and $rank(I - \frac{1}{n}J) = n - 1$. □

Remark 12.4.1.

- The matrix $\frac{1}{n}J$ has rank 1 is because it only has one linearly independent column.
- The matrix $I - \frac{1}{n}J$ is because $rank(I - \frac{1}{n}J) \geq rank(I) - rank(\frac{1}{n}J) = n - 1$ [Theorem 4.4.1]. Also $I - \frac{1}{n}J$ has eigenvector 1 associated with eigenvalue 0. Therefore, $rank(I - \frac{1}{n}J) < n$. In summary, we have $rank(I - \frac{1}{n}J) = n - 1$.

- The matrix $I - \frac{1}{n}J$ has rank $n-1$ because it is orthogonal projector($P^T = P, P^2 = P$) and $rank(I - \frac{1}{n}J) = Tr(I - \frac{1}{n}J) = n - 1$. [Theorem 4.5.7]

12.5 Notes on bibliography

For an extensive discussion on statistical distributions, see [1][2].

BIBLIOGRAPHY

1. Forbes, C., Evans, M., Hastings, N. & Peacock, B. *Statistical distributions* (John Wiley & Sons, 2011).
2. Krishnamoorthy, K. *Handbook of statistical distributions with applications* (CRC Press, 2016).
3. Hogg, R. V., McKean, J. & Craig, A. T. Introduction to Mathematical Statistics, 7 ed (2012).
4. Borovkova, S., Permana, F. J. & Weide, H. V. A closed form approach to the valuation and hedging of basket and spread options. *Journal of Derivatives* **14,** 8 (2007).
5. Brzezniak, Z. & Zastawniak, T. *Basic stochastic processes: a course through exercises* (Springer Science & Business Media, 1999).
6. Murphy, K. P. *Machine learning: a probabilistic perspective* (MIT press, 2012).
7. Casella, G. & Berger, R. L. *Statistical inference* (Duxbury Pacific Grove, CA, 2002).
8. Ruppert, D. *Statistics and data analysis for financial engineering, 2ed* (Springer, 2015).

13

STATISTICAL ESTIMATION THEORY

13.1 Parameter estimators

13.1.1 Overview

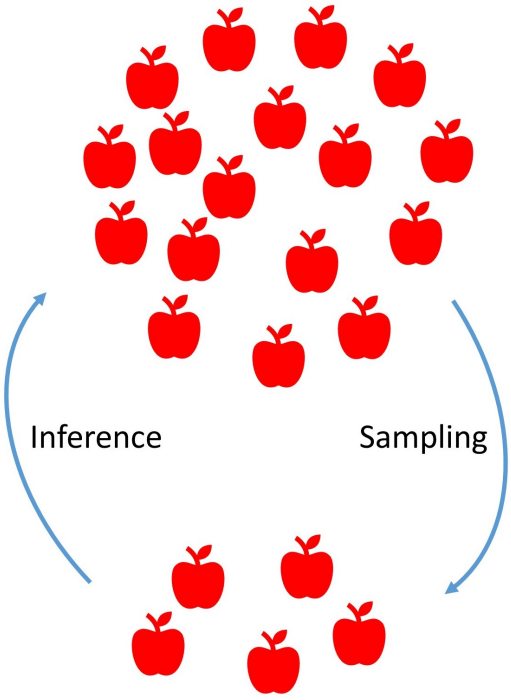


Figure 13.1.1: Statistical estimation and inference scheme.

Given observations of a random variable X, the goal of statistical estimation inference is to infer the population distribution of X from observed samples [Figure 13.1.1]. Direct inference on the population distribution using empirical distribution have several drawbacks. The data requirement usually goes up exponentially with the dimensionality. Further, empirical distributions lack the analytical tractability and convenience if when we need to develop further models based on the empirical distribution.

Instead, we often assume the distribution of X has some parametric form(e.g. Gaussian, Binomial,Poisson). More formally, we assume the distribution of X belongs to a family of distributions for X, $\mathcal{P} = \{P_\theta, \theta \in \Omega\}$, where Ω is the parameter space containing all possible values of θ. Note that a statistical model is a hypothesis, which might be correct or incorrect.

With the statistical model proposed, we estimate the model parameter θ from the data. Once estimated, we have a way to describe the distribution of X, which finishes the inference task.

There are two major components in statistical inference: **proposing statistical models** and **estimating model parameters**.

In proposing statistical models, we consider a number of factors including boundedness of the observed data, mathematically convenience, and tolerance of modeling error. For example, Gaussian distributions have well-established properties and modeling capacity.

In estimating model parameters, we will first design a statistic δ, which is a function of random samples $D = \{X_1, ..., X_N\}$, such that δ is closed to our target θ. In this way, we use a statistic to connect observations D to model parameter θ.

How to design a statistic? It turns out that not all the statistics are created equal. Some are biased and some are more efficient in terms of using observed samples to infer target parameters.

In this chapter, we present foundation and principles in designing good statistics with balanced trade-offs, inferring model parameters, and testing hypothesis regarding statistical models.

13.1.2 Statistic and Estimator

13.1.2.1 *Statistic*

Let $X_1, X_2, ..X_n$ denote random samples from a distribution. Let $T = T(X_1, X_2, ..., X_n)$ be a function of these samples. Then T is called a **statistic**. T is also a random variable.

In the statistical estimation tasks, statistics are design in a way such that their means are model parameters to be estimated. To start with, the most common, and perhaps simplest, statistics are the following. In the subsequent sections, we will discuss a wide variety of statistics specific to different statistical models tasks.

Definition 13.1.1 (common statistic). *Given a random sample $X_1, ..., X_n$ from X, we have following definitions:*

- *Sample mean:*

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

- *Sample variance:*
$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$
- *Sample standard deviation:*
$$S = \sqrt{S^2}.$$

Remark 13.1.1 (another equivalent form of sample variance). Note that $\sum_{i=1}^{n}(X_i - \overline{X})^2$ can also be written by $\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i - X_j)^2$. We have

$$\begin{aligned}\sum_{i=1}^{n}(X_i - \overline{X})^2 &= \sum_{i=1}^{n}(X_i^2 - 2\overline{X}X_i + \overline{X}^2)\\ &= \sum_{i=1}^{n} X_i^2 - 2\overline{X}n\overline{X} + n\overline{X}\\ &= \sum_{i=1}^{n} X_i^2 - n\overline{X}\overline{X},\end{aligned}$$

and

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i - X_j)^2 &= \sum_{i=1}^{n} 2nX_i^2 - \sum_{i=1}^{n}\sum_{j=1}^{n} 2X_iX_j\\ &= \sum_{i=1}^{n} 2nX_i^2 - \sum_{i=1}^{n} 2X_in\overline{X}\\ &= 2n(\sum_{i=1}^{n} X_i^2 - n\overline{X}\overline{X})\end{aligned}$$

13.1.2.2 *Estimator properties*

Given random samples $X_1, ..., X_n$ drawn from a distribution, we often design statistics or estimators $\hat{\theta}(X_1, ..., X_n)$ and use their mean $E[\hat{\theta}]$ as an estimate for parameters θ that describes the distribution.

In terms of estimation quality, the first important characteristic of $\hat{\theta}$, is the estimator's bias. The bias of an estimator $\hat{\theta}$ depicts on average how far $\hat{\theta}$ is from the real value of θ, which we elaborate in the following definition.

Definition 13.1.2 (unbiased estimator). *Let $X_1, X_2, ..X_n$ denote random samples from a distribution. Let $\hat{\theta} = \hat{\theta}(X_1, X_2, ..., X_n)$ be a statistic.*

- *The **bias of an estimator** $\hat{\theta}$ is*

$$Bias(\hat{\theta}) = E[\hat{\theta}] - \theta,$$

  where θ is the true value.
- *If $Bias(\hat{\theta}) = 0$, then estimator $\hat{\theta}$ is said to be **unbiased**, i.e., $E[\hat{\theta}] = \theta$.*

Example 13.1.1 (sample mean estimator is unbiased). The sample mean estimator

$$\hat{\theta} = \overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

is an unbiased estimator of $\theta = E[X_i]$.

Besides the mean, we also need the variance of the estimator as another metric to gauge the quality of an estimator. This is motivated by the following example: In estimating population mean, we can also choose $\hat{\theta} = X_1$, then $\hat{\theta}$, which is an unbiased estimator of θ as well.

However, $\overline{X}$ has a smaller variance than X_1 [a direct result from law of large numbers, Theorem 11.11.1, meaning that $\overline{X}$, as a random variable, has a higher probability of being closer to the true value.

Definition 13.1.3 (variance of an estimator). *The variance of an estimator is defined by*

$$Var[\hat{\theta}] = E[(\hat{\theta} - \theta)^2].$$

*If $\hat{\theta}$ is an **unbiased** estimator, then*

$$Var[\hat{\theta}] = E[\hat{\theta}^2] - \theta^2.$$

Example 13.1.2. Continue the mean estimation example. We assume $Var[X_1] = \sigma^2$. The variance of the estimator $\overline{X}$ is given by

$$Var[\overline{X}] = \sigma^2/n \leq Var[X_1].$$

Example 13.1.3. Suppose that $Y_1, \ldots, Y_n$ are a random sample from a Uniform $(0, \theta)$ distribution where $\theta > 0$ is a parameter. Consider the estimator

$$\hat{\theta}_1 = 2\overline{Y}.$$

We can show that $E[2\overline{Y}] = 2 \cdot \frac{\theta}{2} = \theta$, $Var[Y_i] = \theta^2/12$ for all i, and

$$Var[\overline{\theta}] = Var[2\overline{Y}] = \frac{4}{n} Var[Y_1] = \frac{\theta^2}{3n}.$$

The mean and variance metric further can be unified using mean squared error (MSE) in the following. In general, an estimators will smaller MSE are preferred.

Definition 13.1.4 (mean squared error of an estimator). *The mean squared error(MSE) of an estimator is defined by*

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2].$$

13.1.2.3 *Variance-bias decomposition*

In previous section, we introduce MSE to measure the average squared difference between the estimator $\hat{\theta}$ and the parameter θ. Apart from its simplicity, MSE also has critical connections with bias and variance.

Theorem 13.1.1 (variance bias decomposition). *The MSE of an estimator is related to its variance and bias via*

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = Var[\hat{\theta}] + (Bias(\hat{\theta}))^2$$

where $Var[\hat{\theta}] = E[(\hat{\theta} - E[\hat{\theta}])^2]$. Particularly, if the estimator is unbiased (i.e. $Bias(\hat{\theta}) = 0$), we have

$$MSE(\hat{\theta}) = Var[\hat{\theta}]$$

Proof. Make $(\hat{\theta} - \theta) = (\hat{\theta} - E[\hat{\theta}] + E[\hat{\theta}] - \theta)$ and note that $\hat{\theta}$ is a random variable. Specifically,

$$\begin{aligned}
MSE(\hat{\theta}) &= E[(\hat{\theta} - \theta)^2] \\
&= E[((\hat{\theta} - E[\hat{\theta}] + E[\hat{\theta}] - \theta))^2] \\
&= E[((\hat{\theta} - E[\hat{\theta}])^2] + 2E[((\hat{\theta} - E[\hat{\theta}])(E[\hat{\theta}] - \theta))] + E[(E[\hat{\theta}] - \theta))^2] \\
&= Var[\hat{\theta}] + 2E[((\hat{\theta} - E[\hat{\theta}])](E[\hat{\theta}] - \theta)) + Bias[\hat{\theta}]^2 \\
&= Var[\hat{\theta}] + 0 + Bias[\hat{\theta}]^2
\end{aligned}$$

□

Remark 13.1.2 (bias can be useful). At first glance, it may seem that bias is always undesired. However, biased estimator might have smaller variance [Figure 13.1.2]. As a consequence, biased estimator can have smaller MSE than unbiased estimator. Also consider the following example.

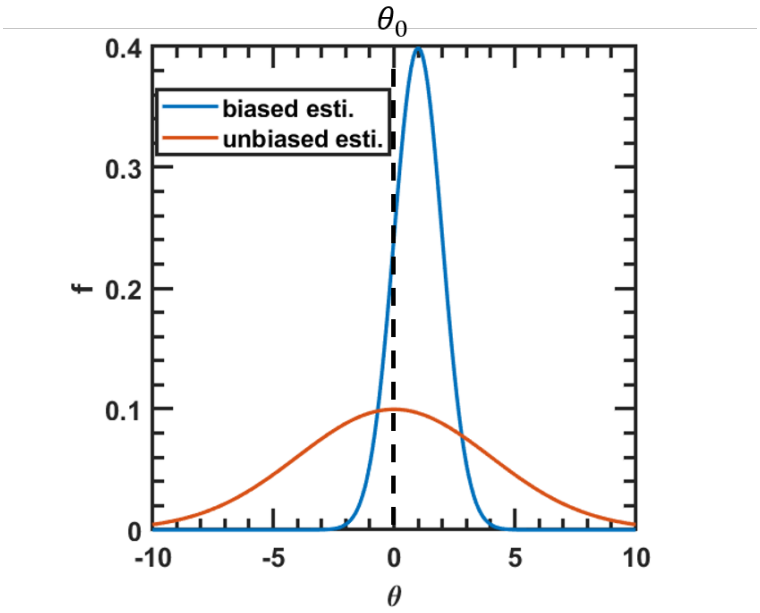


Figure 13.1.2: An example of biased estimator with smaller variance than unbiased estimator

Example 13.1.4. Consider a sample $X_1, X_2, ..., X_n$ of iid normal random variable with unknown mean and variance. Consider two variance estimator

$$S_1^2 = \frac{\sum_{i=1}^n (X_i - \overline{X})}{n-1}, S_2^2 = \frac{\sum_{i=1}^n (X_i - \overline{X})}{n}.$$

Then

- The MSE for S_1^2 is

$$\begin{aligned} MSE[S_1^2] &= Var[S_1^2] + [Bias]^2 \\ &= \frac{2\sigma^4}{n-1} + 0 = \frac{2\sigma^4}{n-1} \end{aligned}$$

  where we use the fact that

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1),$$

from Theorem 12.4.2 and

$$Var[\frac{(n-1)S^2}{\sigma^2}] = \frac{(n-1)^2}{\sigma^4} Var[S^2] = 2(n-1),$$

where use $Var[\chi^2(n)] = 2n$ in Lemma 12.1.31.

- The MSE for S_2^2 is

$$\begin{aligned} MSE[S_2^2] &= (Var[S_2^2] + [Bias])^2 \\ &= \frac{2(n-1)\sigma^4}{n} + (E[S_2^2] - \sigma^2)^2 \\ &= \frac{2(n-1)\sigma^4}{n} + (\frac{(n-1)\sigma^2}{n} - \sigma^2)^2 \\ &= \frac{2n-1}{n^2}\sigma^4 \end{aligned}$$

- $MSE[S_2^2] < MSE[S_1^2]$. That is, the maximum-likelihood estimator has smaller MSE than the unbiased estimator.

13.1.2.4 *Consistence*

We also like to investigate the behavior of estimators as the sample size n gets larger. Understanding of this behavior can be particularly useful in designing experiments involving large scale data in the big data era, We say an estimator is consistent if $\hat{\theta}$ converges to the real value θ when sample size approaches infinity. More precisely, we have the following definition.

Definition 13.1.5 (consistent estimator). *We say $\hat{\theta}$ is a **consistent** estimator of θ if $\hat{\theta}$ converges to θ in probability, i.e.,*

$$\lim_{n\to\infty} P(\left|\hat{\theta}(X_1, X_2, ..., X_n) - \theta\right| < \epsilon) = 1, \forall \epsilon > 0.$$

Example 13.1.5 (sample mean estimator is consistent). Let $X_1, X_2, X_3, \ldots, X_n$ be random samples with the same mean θ, and variance σ^2. We can use Chebyshev's inequality [Theorem 11.9.1] to write

$$P(|\bar{X} - \theta| \geq \epsilon) \leq \frac{Var[\bar{X}]}{\epsilon^2}$$
$$= \frac{\sigma^2}{n\epsilon^2}$$

which goes to 0 as $n \to \infty$.

A more convenient criterion to check if an estimator is consistent is via the following MSE criterion.

Theorem 13.1.2 (MSE criterion for consistent estimator). *An estimator $\hat{\theta}$ is consistent if*

$$\lim_{n\to\infty} MSE(\hat{\theta}(X_1, X_2, ..., X_n)) = 0,$$

*In particular, an **unbiased** estimator $\hat{\theta}$ is consistent if*

$$\lim_{n\to\infty} Var(\hat{\theta}(X_1, X_2, ..., X_n)) = 0,$$

Proof. Overall, we can use Theorem 11.10.3 (convergence in mean square implies convergence in probability). Specifically, we have

$$P\left(\left|\hat{\theta}_n - \theta\right| \geq \epsilon\right) = P\left(\left|\hat{\theta}_n - \theta\right|^2 \geq \epsilon^2\right)$$
$$\leq \frac{E\left[\hat{\theta}_n - \theta\right]^2}{\epsilon^2} \quad \text{(by Markov's inequality)}$$
$$= \frac{MSE\left(\hat{\theta}_n\right)}{\epsilon^2}$$

which goes to 0 as $n \to \infty$ and $MSE \to 0$ by the assumption. □

Remark 13.1.3 (consistence vs. bias)**.**

- A consistent estimator is at least **asymptotically unbiased**. However, some unbiased estimators can be inconsistent (i.e. the variance does not converge to 0), say X_1 as the mean estimator.
- If the sample size is large, consistent estimators are considered better than unbiased estimators because consistent estimators ensure that estimator variance goes to sufficiently smaller when sample size is large.

- Inconsistent estimators usually should be avoided, since increasing the number of samples will not necessarily reduce the variance.

13.1.2.5 *Efficiency*

We now introduce the concept of **efficiency** of an estimator. Essentially, we characterize an efficient estimator by the fact that **given a fixed number of samples**, a more efficient estimator has a lower MSE/variance. Further, the **relative efficiency** of two **unbiased** estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ is the ratio of their variance

$$\frac{Var[\hat{\theta}_1]}{Var[\hat{\theta}_2]}.$$

Example 13.1.6 (sample mean is the most efficient linear estimator of population mean). Consider the linear estimator

$$\widehat{\theta}_n = \sum_{i=1}^{n} a_i X_i$$

where $E[X_i] = \theta, Var[X_i] = \sigma^2$ for $1 \leq i \leq n$.

$$E\left[\widehat{\theta}_n\right] = \sum_{i=1}^{n} a_i E[X_i] = \theta \sum_{i=1}^{n} a_i,$$

so the estimator is unbiased provided

$$\sum_{i=1}^{n} a_i = 1.$$

For i.i.d. random variables,

$$Var\left[\widehat{\theta}_n\right] = \sum_{i=1}^{n} a_i^2 \sigma^2.$$

Now from constrained optimization theory, we know that minimum is achieved iff $a_i = \frac{1}{n}$, $1 \leq i \leq n$. The conclusion then is that, if $\widehat{\theta}_n$ is a linear unbiased estimator of the form $\sum_{i=1}^{n} a_i X_i$ and if $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$, then

$$Var\left[\overline{X}\right] \leq Var\left[\widehat{\theta}_n\right].$$

Therefore, among all linear unbiased estimators, $\overline{X}$ *is most efficient* estimator of the population mean.

Remark 13.1.4 (unbiasedness and efficiency). A biased estimator with a small variance may be more useful than an unbiased estimator with a large variance.

13.1.2.6 *Robust statistics*

Robust statistics are statistics that resilient to sample outliers, samples that are roughly significantly distinct from the majority of the samples. To pave the way for the discussion on robust statistics, we first introduce the concept of **breakdown point** for an estimator. The finite sample **breakdown point** of an estimator is the smallest fraction α of data points such that if $[n\alpha]$ points approach ∞, then the estimator approach ∞.

Given sample size n, the breakdown point for sample mean estimator using the arithmetic mean is $1/n$; that is one point can ruin the mean. To see this, we have

$$\begin{aligned}\overline{x}_n &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ &= \frac{1}{n}(\sum_{i=1}^{n} x_i + x_n) \\ &= \frac{n-1}{n}(\overline{x}_{n-1}) + \frac{1}{n}x_n\end{aligned}$$

Suppose $x_1, ..., x_{n-1}$ are well-behaving points. When x_n, as one outlier, approaches to ∞, $\overline{x}_n$ will approach ∞. As a comparison, the sample median, as an estimate of a population median, can tolerate up to 50% bad values, i.e., its breakdown point is 0.5.

In the following, we introduce some robust estimators for mean and variance.

Definition 13.1.6 (α trimmed mean). *Let $k = n\alpha$ rounded to an integer(k is the number of observation removed from both ends for calculation). The* ***α-trimmed mean*** *is defined as*

$$\overline{X}_\alpha = \sum_{i=k+1}^{n-k} \frac{X_i}{n-2k}.$$

Definition 13.1.7 (median absolute deviation). *[1, p. 122] A robust estimator of standard deviation of iid random sample $X_1, X_2, ..., X_n$ is the* ***MAD(median absolute deviation)***

$$\hat{\sigma}^{MAD} = 1.4826 \times median\{|X_i - median(X_i)|\}.$$

Remark 13.1.5 (interpretation).

- For normally distributed data, $median\{|X_i - median(Y_i)|\}$ is the estimator of $\Phi^{-1}(0.75)\sigma = \sigma/1.4826$.
- For an iid normal random sample, as sample size $n \to \infty$, the MAD is the unbiased estimates of σ.

13.1.3 Method of moments

The method of moments is a straight forward approach to set up the equation to find estimators, although the quality of the found estimators can be of low quality and sovling equations can be troublesome in some cases.

To start with, let $X_1, X_2, ..., X_n$ be a sample from a population with pdf or pmf $f(x|\theta_1, ..., \theta_k)$. The first k moments of the random samples are given by

$$\begin{aligned} m_1 &= \frac{1}{n}\sum_{i=1}^{n} X_i \\ m_2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 \\ &\cdots \\ m_k &= \frac{1}{n}\sum_{i=1}^{n} X_i^k \end{aligned}$$

From the assumed pdf or pmf $f_X(x;\theta)$ of a random sample, we can derive the theoretical moments as a function of parameter θ. By equating theoretical moments and sample moment, we can solve θ. We elaborate this approach in the following.

Methodology 13.1.1 (method of moments for parameter estimation). *[2, p. 312] Let $X_1, X_2, ..., X_n$ be a random sample of X from a population with pdf or pmf $f(x|\theta_1, ..., \theta_k)$.*

Define $\mu_i = E[X^i], i = 1, 2, ..., k$. *The method of moments is aimed at solving* $\theta_1, \theta_2, ..., \theta_k$ *from the k equations*

$$\begin{aligned} m_1 &= \mu_1(\theta_1, ..., \theta_k) \\ m_2 &= \mu_2(\theta_1, ..., \theta_k) \\ &\cdots \\ m_k &= \mu_k(\theta_1, ..., \theta_k) \end{aligned}$$

where $m_i, i = 1, ..., k$ *are the k sample moments, and* $\mu_i, i = 1, .., k$ *are theoretical samples given by*

$$\mu_i E[X^i] = \int x^i f(x) dx.$$

Example 13.1.7 (estimating normal distribution parameter via method of moments). Suppose $X_1, X_2, ..., X_n$ are iid random samples with distribution $N(\mu, \sigma^2)$. It follows that

- Theoretical moments are $\mu_1 = \mu, \mu^2 + \sigma^2 = \mu_2$.
- The moment of method estimators for (μ, σ) are

$$\hat{\mu} = m_1, \hat{\sigma}^2 = m_2 - m_1^2.$$

Example 13.1.8 (estimating t distribution parameter via method of moments). Suppose $X_1, X_2, ..., X_n$ are iid random variable with $t_v(\mu, \sigma^2), v > 2$. It follows that

- Theoretical moments are $m_1 = \mu, \mu^2 + \sigma^2 \frac{v}{v-2} = m_2$.
- The moment of method estimators for (μ, σ) are

$$\hat{\mu} = m_1, \hat{\sigma}^2 = (m_2 - m_1^2) \frac{v-2}{v}.$$

13.1.4 Maximum likelihood estimation

13.1.4.1 *Basic concepts*

The method of moments is quite dated. Nowadays, the arguably most popular point estimation method is maximum likelihood estimation (MLE). The strengths of MLE include its simplicity, generality and efficiency [subsection 13.2.4]. Its dominance is further promoted by recent progress of numerical software and technology (e.g., automatic

differentiation), which make MLE a viable computational tool for large-scale, complex statistical modeling problems.

In general, Given observations $\boldsymbol{x} = X_1, X_2, ..., X_n$ random samples, MLE is comprised of two steps

- Find the likelihood function $L(\boldsymbol{x}|\theta)$ based on observed samples and assumed parametric distribution. The likelihood function is a function of the parameter θ.
- Find the optimal θ that maximizes the likelihood function L. θ^* is the MLE.

Definition 13.1.8 (likelihood function and MLE). *Assuming a statistical model parameterized by a fixed and unknown θ, the likelihood $L(\boldsymbol{x}|\theta)$ is the probability of the observations $\boldsymbol{x} = (x_1, x_2, ..., x_n)$ of iid random samples $X_1, X_2, ..., X_n$ as a function of θ. It can be written as*

$$L(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} f(X = x_i|\theta)$$

And the corresponding log-likelihood function is defined by

$$\log L(\boldsymbol{x}|\theta) = \sum_{i=1}^{n} f(X = x_i|\theta).$$

A maximum likelihood estimator(MLE) of the parameter θ based on given observations $\boldsymbol{x}$ is

$$\hat{\theta} = \max_{\theta} \log L(\boldsymbol{x}|\theta),$$

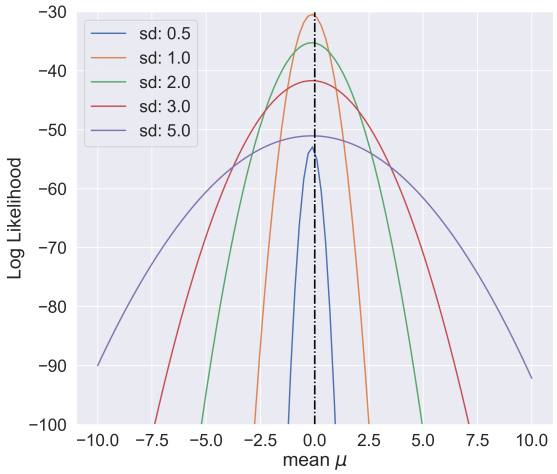
Or alternatively, $\hat{\theta}$ satisfies

$$s(\theta, \boldsymbol{x}) = \frac{\partial \log L}{\partial \theta} = 0,$$

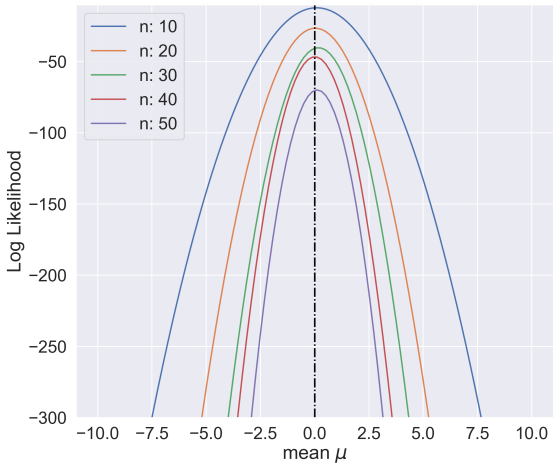
*where $s(\theta, \boldsymbol{x})$ is called **score function**.*

In chapter 12, we examine the log-likelihood functions for samples drawn from a normal distribution $N(0, 1)$. We assume the population distribution is normal governed by parameters μ and σ.

For a set of fixed samples, the log-likelihood function varies with μ and σ [Figure 13.1.3(a)]. Clearly, log-likelihoods achieve peak values when μ is **near** the true value. Log-likelihood is also a function of the sample size n [[Figure 13.1.3(b)]]. For large sample size, log-likelihood will be more sensitive to the parameters, that is, changing more rapidly when parameters change.



(a) Log-likelihood function as a function of μ and θ. Sample size 20.



(b) Log-likelihood function as a function of μ and sample size n. $\sigma = 1$.

Figure 13.1.3: Visualization of log-likelihood function for normal distributed samples.

13.1.4.2 *MLE examples*

Example 13.1.9 (Normal distribution MLE). The log-likelihood function for n iid observations $x_1, ..., x_n$ drawn from normal distribution is given by

$$\begin{aligned}\log L(\theta_1, \theta_2) &= \prod_{i=1}^{n} f(x_i \theta_1, \theta_2) \\ &= \theta_2^{-n/2}(2\pi)^{-n/2} \exp(-\frac{1}{2\theta_2}\sum_{i=1}^{n}(x_i - \theta_1)^2) \\ &= -\frac{n}{2}\log\theta_2 - \frac{n}{2}\log(2\pi) - \frac{\sum_{i=1}^{n}(x_i - \theta_1)^2}{2\theta_2}\end{aligned}$$

where $\theta_1 = \mu, \theta_2 = \sigma^2$. Setting derivatives to zeros, we have

$$\begin{aligned}\frac{\partial \log L(\theta_1, \theta_2)}{\partial \theta_1} &= \frac{\sum_{i=1}^{n}(x_i - \theta_1)}{\theta_2} = 0 \\ \frac{\partial \log L(\theta_1, \theta_2)}{\partial \theta_2} &= -\frac{n}{2\theta_2} + \frac{\sum_{i=1}^{n}(x_i - \theta_1)^2}{\theta_2^2} = 0\end{aligned}$$

which produces

$$\hat{\mu} = \hat{\theta}_1 = \frac{\sum_{i=1}^{n} x_i}{n} = \overline{x}, \hat{\sigma}^2 = \hat{\theta}_2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x}}{n}.$$

Example 13.1.10 (Bernoulli trial MLE). Consider a series of independent Bernoulli trials with success probability θ such that we have probability mass function given by

$$Pr(Y_i = y) = (1-\theta)^{1-y}\theta^y, y \in \{0, 1\}.$$

- The log-likelihood function based on n observations $Y = \{Y_1, ..., Y_N\}$ can be written by

  $$\log L(\theta; Y) = \sum_{i=1}^{n}((1-y_i)\log(1-\theta) + y_i \log\theta) = n((1-\overline{y})\log(1-\theta) + \overline{y}\log(\theta)),$$

  where $\overline{y}$ is the sample mean.
- The MLE is given by

  $$\hat{\theta} = \overline{y}.$$

Example 13.1.11 (exponential distribution MLE). Consider an exponential distribution with parameter α such that its pdf is given by

$$f(x;\alpha) = \alpha e^{-\alpha x}, x \geq 0.$$

The MLE for α from an iid random sample $X_1, ..., X_n$ is given by $\hat{\alpha} = 1/\overline{X}$ since

$$\log L(\alpha) = n \log \alpha - \alpha \sum_{i=1}^{n} X_i$$
$$\partial \log L(\alpha)/\partial \alpha = \frac{n}{\alpha} - \sum_{i=1}^{n} X_i$$
$$\partial \log L(\alpha)/\partial \alpha = 0 \implies \hat{\alpha} = 1/\overline{X}.$$

Example 13.1.12. Consider an exponential distribution with parameter α such that its pdf is given by

$$f(x;\alpha) = \alpha e^{-\alpha x}, x \geq 0.$$

- The MLE for α from an iid random sample $X_1, ..., X_n$ is given by $\hat{\alpha} = 1/\overline{X}$ since

$$\log L(\alpha) = n \log \alpha - \alpha \sum_{i=1}^{n} X_i$$
$$\partial \log L(\alpha)/\partial \alpha = \frac{n}{\alpha} - \sum_{i=1}^{n} X_i$$
$$\partial \log L(\alpha)/\partial \alpha = 0 \implies \hat{\alpha} = 1/\overline{X}.$$

- The Fisher information is given by $I(\alpha) = \frac{1}{\alpha^2}$ since

$$\log f(x;\alpha) = \log \alpha - \alpha x$$
$$\partial^2 \log L(\alpha)/\partial \alpha^2 = -\frac{1}{\alpha^2}$$
$$I(\alpha) = -E[\partial^2 \log L(\alpha)/\partial \alpha^2] = \frac{1}{\alpha^2}.$$

- The MLE $\hat{\alpha} = 1/\overline{X}$ is asymptotic normal and

$$\sqrt{n}(\hat{\alpha} - \alpha_0) \to N(0, \alpha_0^2).$$

Example 13.1.13. Let $X_1, X_2, X_3, \ldots, X_n$ be a random sample from a U niform $(0, \theta)$ distribution, where θ is unknown. Find the maximum likelihood estimator (MLE) of θ based on this random sample. Solution If $X_i \sim$Uniform$(0, \theta)$, then

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

The likelihood function is given by

$$\begin{aligned} L(x_1, x_2, \cdots, x_n; \theta) &= f_{X_1 X_2 \cdots X_n}(x_1, x_2, \cdots, x_n; \theta) \\ &= f_{X_1}(x_1; \theta) f_{X_2}(x_2; \theta) \cdots f_{X_n}(x_n; \theta) \\ &= \begin{cases} \frac{1}{\theta^n} & 0 \leq x_1, x_2, \cdots, x_n \leq \theta \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Note that $\frac{1}{\theta^n}$ is a decreasing function of θ. Thus, to maximize it, we need to choose the smallest possible value for θ. For $i = 1, 2, \ldots, n$, we need to have $\theta \geq x_i$. Thus, the smallest possible value for θ is

$$\hat{\theta}_{ML} = \max(x_1, x_2, \cdots, x_n)$$

Therefore, the MLE can be written as

$$\hat{\theta}_{ML} = \max(X_1, X_2, \cdots, X_n)$$

Note that this is one of those cases wherein $\hat{\theta}_{ML}$ cannot be obtained by setting the derivative of the likelihood function to zero. Here, the maximum is achieved at an endpoint of the acceptable interval.

13.1.4.3 *Bias and consistence of MLE*

An MLE could be biased. For example, the ML variance estimator for normal samples $X_1, X_2, ..., X_n$ is

$$S^2_{ML} = \frac{\sum_{i=1}^n (X_i - \overline{X})}{n},$$

which is biased, as opposed to the unbiased sample variance estimator

$$S^2 = \frac{\sum_{i=1}^n (X_i - \overline{X})}{n-1}.$$

However, the ML variance estimator has a smaller variance, as showed in the following example.

Example 13.1.14 (ML variance estimator has a smaller variance). Consider samples $X_1, X_2, ..., X_n$ of iid normal random variable with unknown mean and variance. Consider two variance estimator

$$S_1^2 = \frac{\sum_{i=1}^n (X_i - \overline{X})}{n-1}, S_2^2 = \frac{\sum_{i=1}^n (X_i - \overline{X})}{n}.$$

Then

- The MSE for S_1^2 is

$$\begin{aligned} MSE[S_1^2] &= Var[S_1^2] + [Bias]^2 \\ &= \frac{2\sigma^4}{n-1} + 0 = \frac{2\sigma^4}{n-1} \end{aligned}$$

  where we use the fact that

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1),$$

  from Theorem 12.4.2 and

$$Var[\frac{(n-1)S^2}{\sigma^2}] = \frac{(n-1)^2}{\sigma^4} Var[S^2] = 2(n-1),$$

  where use $Var[\chi^2(n)] = 2n$ in Lemma 12.1.31.
- The MSE for S_2^2 is

$$\begin{aligned} MSE[S_2^2] &= (Var[S_2^2] + [Bias])^2 \\ &= \frac{2(n-1)\sigma^4}{n} + (E[S_2^2] - \sigma^2)^2 \\ &= \frac{2(n-1)\sigma^4}{n} + (\frac{(n-1)\sigma^2}{n} - \sigma^2)^2 \\ &= \frac{2n-1}{n^2}\sigma^4 \end{aligned}$$

- $MSE[S_2^2] < MSE[S_1^2]$. That is, the maximum-likelihood estimator has smaller MSE than the unbiased estimator.

Despite its possible bias for finite samples, one important property of MLEs is that as sample size $n \to \infty$, they would converge to the true value. We say that they are asymptotically unbiased, or consistent.

Theorem 13.1.3 (consistence of MLE). *Let $\hat{\theta}$ be the MLE of coefficient associated with distribution $f(x;\theta)$. Let θ_0 be the true value of the parameter. Let $I_1(\theta)$ be the Fisher information matrix associated with distribution $f(x;\theta)$. It follows that MLEs are consistent; that is*

$$\operatorname*{plim}_{n\to\infty} \hat{\theta} = \theta_0.$$

Proof. (sketch) Define the log-likelihood function associated with n samples

$$L_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log f(X_i|\theta).$$

Denote MLE $\hat{\theta}_n = \arg\max_\theta L_n(\theta)$. Consider the function $L(\theta) = \int (\log f(x;\theta) f(x;\theta_0) dx$, now we can show that the **true parameter θ_0 is the maximizer of** $L(\theta)$; that is, for any θ, we have

$$L(\theta) \le L(\theta_0).$$

$$\begin{aligned}
L(\theta) - L(\theta_0) &= E_{\theta_0}[\log f(X;\theta) - \log f(X;\theta_0)] \\
&= E_{\theta_0}[\log \frac{f(X;\theta)}{f(X;\theta_0)}] \\
&\le E_{\theta_0}[\frac{f(X;\theta)}{f(X;\theta_0)} - 1] \\
&= \int (\frac{f(X;\theta)}{f(X;\theta_0)} - 1) f(x;\theta_0) dx \\
&= \int f(x;\theta) dx - \int f(x;\theta_0) dx \\
&= 1 - 1 = 0
\end{aligned}$$

where we use inequality $\log x \le x - 1$.

From the law of large numbers, $L_n(\theta)$ converge to $L(\theta)$ in probability. Since MLE $\hat{\theta}$ is the maximizer for $L_n(\theta)$, $\hat{\theta}$ converges to θ_0 in probability. □

13.2 Information and efficiency

13.2.1 Fisher information

One central task of statistical estimation is to construct efficient estimators that can extract most information from finite number of samples. In previous section [subsubsection 13.1.2.5], efficiency of an unbiased estimator is characterized by its variance. In this section, we discuss Fisher information framework, which provides a lower bound on the variance of estimators given finite number of samples. The knowledge of lower bound enables us to assert if the estimator we found has the lowest possible variance.

To start with, we discuss Fisher information definitions and its important properties; in the subsequent sections, we examine how Fisher information is utilized to derive the lower bound on variances.

Fisher information requires some reguarlity conditions on the pdf for continuous random variables.

Assumption 13.1 (Fisher information regularity assumption). *For a pdf $f(x;\theta)$ of random variable X with parameter θ. We make the following regularity assumptions:*

- *The set $A = \{x|p(x;\theta) > 0\}$ does not depend on θ. For all $x \in A, \theta \in \Theta$, $\frac{\partial}{\partial\theta}\log p(x;\theta)$ exists and is finite. Here Θ is the parameter space.*
- *If T is any statistic of X such that $E[|T|] < \infty$ for all $\theta \in \Theta$, then integration and differentiation by θ can be interchanged in the following way:*

$$\frac{\partial}{\partial\theta}[\int T(x)f(x;\theta)dx] = \int T(x)\frac{\partial}{\partial\theta}f(x;\theta)dx,$$

whenever the right-hand side is finite.

The definition of Fisher information for discrete or continuous random variables is given by the following.

Definition 13.2.1 (Fisher information). *Given one dimensional parametric family of pdf or pmf $f(x;\theta)$, which is differentiable respect to θ, we define the Fisher information for $\theta \in \mathbb{R}$ as*

$$I(\theta) = E[(\frac{d}{d\theta}\log f(x;\theta))^2],$$

where the expectation is ***taken with respect to*** x*. In particular, if* $\theta \in \mathbb{R}^N$*, we have Fisher information matrix defined by*

$$I(\theta) = E[\frac{\partial \log(f(x;\theta))}{\partial \theta}(\frac{\partial \log(f(x;\theta))}{\partial \theta})^T].$$

The derivative $\frac{d}{d\theta}\log f(x|\theta)$ is known as the **score function**, which characterize the sensitivity of f with respect to θ at a particular θ. Then intuitively, the Fisher information measures the overall on average sensitivity.

How the sensitivity will affect the variance of the estimator? Suppose Fisher information is large, so the parametric distribution will change rapidly when parameters change and be quite different from the true distribution. The large difference can be particularly useful in designing numerical algorithms to search for the true parameters. Conversely, if the Fisher information is small and the distribution is insensitive to model parameters, it would be difficult to estimate the true parameters.

Theorem 13.2.1 (basic properties of Fisher information). *Let* $f(x;\theta)$ *be a pdf parameterized by* $\theta \in \mathbb{R}$ *with Assumption 13.1 holds, then*

- *Score function is zero mean:*

$$E[\frac{d}{d\theta}\log f(x;\theta)] = 0$$

- *Fisher information is the variance of the score function:*

$$I(\theta) = Var[\frac{d}{d\theta}\log f(x;\theta)].$$

- *Further assume* $f(x;\theta)$ *is twice differentiable and interchange between integration and differentiation is permitted. Then*

$$I(\theta) = E[(\frac{d}{d\theta}\log f(x;\theta))^2] = -E[\frac{\partial^2 \log(f(x;\theta))}{\partial \theta^2}].$$

- *For* $\theta \in \mathbb{R}^N$*,*

$$I(\theta) = E[\frac{\partial \log(f(x;\theta))}{\partial \theta}(\frac{\partial \log(f(x;\theta))}{\partial \theta})^T] = -E[\frac{\partial^2 \log(f(x;\theta))}{\partial \theta \partial \theta^T}],$$

for each entry in the matrix,

$$I(\theta)_{ij} = -E[\frac{\partial^2 \log(f(x;\theta))}{\partial\theta_i \partial\theta_j}].$$

Proof. (1) The equivalence of these two expressions can be showed as:

$$\begin{aligned} E[\frac{d}{d\theta}\log f(x;\theta)] &= \int \frac{1}{f(x;\theta)}\frac{d}{d\theta}f(x;\theta)f(x;\theta)dx \\ &= \int \frac{d}{d\theta}f(x;\theta)dx \\ &= \frac{d}{d\theta}\int f(x;\theta)dx \\ &= \frac{d}{d\theta}1 \\ &= 0 \end{aligned}$$

(2) Based on definition, we have

$$\begin{aligned} Var[\frac{d}{d\theta}\log f(x;\theta)] &= E[(\frac{d}{d\theta}\log f(x;\theta))^2] - E[\frac{d}{d\theta}\log f(x;\theta)]^2 \\ &= E[(\frac{d}{d\theta}\log f(x;\theta))^2] - 0 \\ &= I(\theta) \end{aligned}$$

(3)

$$\begin{aligned} \frac{\partial^2}{\partial\theta^2}\log f(x;\theta) &= \frac{\partial}{\partial\theta}\frac{1}{f(x;\theta)}\frac{\partial}{\partial\theta}f(x;\theta) \\ &= -\frac{\partial}{\partial\theta}\frac{1}{f(x;\theta)^2}\frac{\partial}{\partial\theta}f(x;\theta) + \frac{1}{f(x;\theta)}\frac{\partial^2}{\partial\theta^2}f(x;\theta) \\ &= -(\frac{\partial}{\partial\theta}\log f(x;\theta))^2 + \frac{1}{f(x;\theta)}\frac{\partial^2}{\partial\theta^2}f(x;\theta) \end{aligned}$$

Take expectation with respect to x on both sides and note that

$$E[\frac{1}{f(x;\theta)}\frac{\partial^2}{\partial\theta^2}f(x;\theta)] = \int \frac{\partial^2}{\partial\theta^2}f(x;\theta)dx = \frac{\partial^2}{\partial\theta^2}\int f(x;\theta)dx = \frac{\partial^2}{\partial^2\theta}1 = 0.$$

□

Since our ultimate goal is to estimate parameters from a set of random samples, it is beneficial to distinguish between Fisher information associated with the distribution of

a random variable and Fisher information associated with the joint distribution of a set of iid random samples. Let X denote a single random variable, and let X^n denote N iid random samples. Then based on iid assumption, we have

$$I_{X^n}(\theta) = nI_X(\theta),$$

where we use the fact that $f_{X^n} = [f_X]^n$. For simplicity, we might also use $I_1(\theta)$ and $I_n(\theta)$ to distinguish them in the following sections.

Example 13.2.1 (Fisher information for Bernoulli distribution). Let the pmf of Bernoulli distribution be $f(x;\theta) = \theta^x(1-\theta)^{1-x}, x \in \{0,1\}$. Then

$$I(\theta) = \frac{1}{\theta(1-\theta)}.$$

To see this, we have

$$\begin{aligned} I(\theta) &= -E[\frac{\partial^2 \log(f(x;\theta))}{\partial\theta^2}] \\ &= E[\frac{x}{\theta^2} + \frac{1-x}{(1-\theta)^2}] \\ &= \theta(1/\theta^2) + (1-\theta)(1/(1-\theta)^2) \\ &= \frac{1}{\theta(1-\theta)} \end{aligned}$$

Example 13.2.2 (Fish information matrix for univariate normal distribution). Let the pdf of normal distribution parameterized by

$$f(x;\theta) = (2\pi\theta_2)^{-1/2}\exp(-\frac{1}{2\theta_2}\sum_{i=1}^{n}(x-\theta_1)^2).$$

$$\begin{aligned} \frac{\partial^2 \log f}{\partial\theta_1^2} &= -\frac{1}{\theta_2} = -\frac{1}{\sigma^2} \\ \frac{\partial^2 \log f}{\partial\theta_2^2} &= \frac{1}{2\theta_2^2} - \frac{1}{\theta_2^3}(x-\theta_1)^2 \\ \frac{\partial^2 \ln f}{\partial\theta_1\partial\theta_2} &= -\frac{1}{\theta_2^2}(x_i-\theta_1) \end{aligned}$$

Finally, take expectation with respect to x and we have

$$I(\theta_1, \theta_2) = \begin{bmatrix} 1/\sigma^2 & 0 \\ 0 & 1/(2\sigma^4) \end{bmatrix}.$$

13.2.2 Cramer-Rao lower bound

13.2.2.1 *Preliminary: information inequality*

Theorem 13.2.2 (information inequality for a statistic). *Let $T(X)$ be any statistic such that $Var[T(X)] < \infty$ for all θ. Denote $E[T(X)]$ by $\phi(\theta)$. Suppose Assumption 13.1 holds and $0 < I(\theta) < \infty$. Then **for all** θ*

$$Var[T(X)] \geq \frac{[\phi'(\theta)]^2}{I(\theta)}.$$

Proof. Based on the Assumption 13.1, ,we have

$$\phi'(\theta) = \int T(x) \frac{\partial}{\partial \theta} f(x;\theta) dx = \int T(x) \frac{\partial \log f(x;\theta)}{\partial \theta} f(x;\theta) dx.$$

Therefore, we can view

$$\phi'(\theta) = E[T(X) \frac{\partial \log f(x;\theta)}{\partial \theta}] = Cov[T(X), \frac{\partial \log f(x;\theta)}{\partial \theta}],$$

since $E[\frac{\partial \log f(x;\theta)}{\partial \theta}] = 0 \implies E[T(X)]E[\frac{\partial \log f(x;\theta)}{\partial \theta}] = 0$ [Theorem 13.2.1].

Using Cauchy-Schwartz inequality [Theorem 11.9.4], we have

$$\left|\phi'(\theta)\right|^2 = Cov[T(X), \frac{\partial \log f(x;\theta)}{\partial \theta}]^2 \leq Var[T(X)] Var[\frac{\partial \log f(x;\theta)}{\partial \theta}].$$

At last, use the fact [Theorem 13.2.1] that $I(\theta) = Var[\frac{\partial \log f(x;\theta)}{\partial \theta}]$, we can get the final result. □

Corollary 13.2.2.1 (information lower bound for general estimators). *Let $T(X)$ be a (generally biased) estimator of θ such that*

$$\phi(\theta) \triangleq E[T(X)] = \theta + \underbrace{b(\theta)}_{bias}.$$

Then

- *the variance of $T(X)$ is*
$$Var[T(X)] \geq \frac{\left|1 + b'(\theta)\right|}{I(\theta)}.$$
- *the MSE of $T(X)$ is*
$$MSE[T(X)] \geq \frac{\left|1 + b'(\theta)\right|}{I(\theta)} + b(\theta)^2.$$

13.2.2.2 *Cramer-Rao lower bound: univariate case*

Theorem 13.2.3 (Cramer-Rao lower bound in univariate estimation). *Let $\hat{\theta}$ be an arbitrary univariate estimator as a function of iid random samples $X_1, ..., X_n$, whose distribution is parameterized by single parameter θ. Let θ_0 be the true value. Then the variance of the estimator $\hat{\theta}$ is bounded by*
$$Var(\hat{\theta}) \geq \frac{(\frac{d}{d\theta} E[\hat{\theta}])^2}{nI_1(\theta_0)},$$
where $I_1(\theta)$ is the Fish information associated with distribution $f(x;\theta)$ and the expectation is taken with respect to x. Particularly, if the estimator $\hat{\theta}$ is unbiased (that is $E[\hat{\theta}] = \theta$), we have
$$Var(\hat{\theta}) \geq \frac{1}{nI_1(\theta_0)}.$$

Proof. Note that the Fisher information $I(\theta)$ associated with the joint distribution of $(X_1, ..., X_n)$ can be expressed by $I(\theta) = nI_1(\theta)$, where $I_1(\theta)$ is the Fisher information associated with $f(x;\theta)$. This is because under iid assumption,
$$E[\log f(x_1, ..., x_n; \theta)] = nE[\log f(x;\theta)].$$
Then use the information inequality [Theorem 13.2.2], we have
$$Var(\hat{\theta}) \geq \frac{(\frac{d}{d\theta} E[\hat{\theta}])^2}{nI_1(\theta_0)}.$$
□

Example 13.2.3 (univariate estimation for normal distributions).

- Consider an unbiased mean estimator $\hat{\mu}$ and an unbiased variance estimator $\hat{\sigma}^2$ for normal distribution with unknown mean μ and variance σ^2. Because

the information matrix is given by Example 13.2.2, the mean estimator has a bounded variance given by

$$Var[\hat{\mu}] \geq \frac{1}{nI_1(\theta)} = \sigma^2/n.$$

- Consider a normal distribution with known mean μ. The variance estimator has a bounded variance given by

$$Var[\hat{\sigma}^2] \geq \frac{1}{nI_1(\theta)} = 2\sigma^4/n.$$

- It is clear that
  - Increasing sample size n will reduce the estimator variance.
  - Mean/variance estimators of random samples drawn from small-variance distributions have inherent smaller variances in their estimators.

13.2.2.3 *Cramer-Rao lower bound: multivariate case*

Theorem 13.2.4 (information inequality for statistic: multivariate case). *Let $T(X)$ be any statistic such that $Var[T(X)] < \infty$ for all θ. Denote $E[T(X)]$ by $\phi(\theta)$. Suppose Assumption 13.1 holds and $0 < I(\theta) < \infty$. Then* ***for all*** *θ*

$$Var[T(X)] \geq [\nabla_\theta \phi]^T [I(\theta)]^{-1} [\nabla_\theta \phi].$$

Proof. Based on the Assumption 13.1, ,we have

$$\phi'(\theta) = \int T(x) \frac{\partial}{\partial \theta} f(x;\theta) dx = \int T(x) \frac{\partial \log f(x;\theta)}{\partial \theta} dx.$$

Therefore, we can view

$$\phi'(\theta) = E[T(X) \frac{\partial \log f(x;\theta)}{\partial \theta}] = Cov[T(X), \frac{\partial \log f(x;\theta)}{\partial \theta}],$$

since $E[\frac{\partial \log f(x;\theta)}{\partial \theta}] = 0 \implies E[T(X)]E[\frac{\partial \log f(x;\theta)}{\partial \theta}] = 0.$

Using Cauchy-Schwartz inequality [Theorem 11.9.4], we have

$$\left|\phi'(\theta)\right|^2 = Cov[T(X), \frac{\partial \log f(x;\theta)}{\partial \theta}]^2 \leq Var[T(X)] Var[\frac{\partial \log f(x;\theta)}{\partial \theta}].$$

At last, use the fact [Theorem 13.2.1] that $I(\theta) = Var[\frac{\partial \log f(x;\theta)}{\partial \theta}]$, we can get the final result. □

Proof. Similar to Theorem 13.2.4, we can show that

$$\frac{\partial \phi(\theta)}{\partial \theta_j} = Cov(T(X), \frac{\partial \log f(x;\theta)}{\partial \theta_j}).$$

For constants $c_1, c_2, ..., c_p$, note that

$$Var[T(X) - \sum_{j=1}^{p} c_j \frac{\partial \log f(x;\theta)}{\partial \theta_j}] = Var[T(X)] + c^T I(\theta) c - 2c^T [\nabla_\theta \phi]$$
$$\geq 0$$

Particularly, the minimum is achieved at $c* = [I(\theta)]^{-1} \nabla_\theta \phi$. Then

$$Var[T(X) - \sum_{j=1}^{p} c_j \frac{\partial \log f(x;\theta)}{\partial \theta_j}] = Var[T(X)] - [\nabla_\theta \phi]^T [I(\theta)]^{-1} [\nabla_\theta \phi] \geq 0.$$

□

Theorem 13.2.5 (Cramer-Rao lower bound in multivariate estimation). *Let $\hat{\theta}$ be a p-dimension **unbiased** estimator as a function of iid random samples $X_1, ..., X_n$, whose distribution is parameterized by parameter vector $\theta \in \mathbb{R}^p, \theta = (\theta_1, ..., \theta_p)$. Let θ_0 be the true value.*

Then the variance matrix of the estimator $\hat{\theta}_i$ is bounded by

$$Var(\hat{\theta}) \geq [n[I_1(\theta_0)]^{-1},$$

where $I_1(\theta_0)$ is the Fish information maxtrix associated with distribution $f(x;\theta)$ and the expectation is taken with respect to x.

Proof. Note that the Fisher information $I(\theta)$ associated with the joint distribution of $(X_1, ..., X_n)$ can be expressed by $I(\theta) = nI_1(\theta)$, where $I_1(\theta)$ is the Fisher information associated with $f(x, \theta)$. This is because under iid assumption,

$$E[\log f(x_1, ..., x_n; \theta)] = nE[\log f(x;\theta)].$$

Then use the information inequality [Theorem 13.2.4], we have

$$Var(\alpha^T \hat{\theta}) = \alpha^T Var[\hat{\theta}] \alpha \geq [\nabla_\theta \alpha^T \hat{\theta}]^T [nI_1(\theta)]^{-1} [\nabla_\theta \alpha^T \hat{\theta}] = \alpha^T [nI_1(\theta)]^{-1} \alpha,$$

where $\alpha \in \mathbb{R}^p$ is an arbitrary vector. □

Example 13.2.4 (multivariate estimation for normal distributions).

- Consider an unbiased mean estimator $\hat{\mu}$ for normal distribution with known variance σ^2. The information matrix is given by (Example 13.2.2)

$$I(\theta_1, \theta_2) = \begin{bmatrix} 1/\sigma^2 & 0 \\ 0 & 1/(2\sigma^4) \end{bmatrix}.$$

  Therefore,

$$Var[\hat{\mu}] \geq \sigma^2/n, Var[\hat{\sigma}^2] \geq 2\sigma^4/n,$$

- It is clear that
  - Increasing sample size n will reduce the estimator variance.
  - Mean/variance estimators of random samples drawn from small-variance distributions have inherent smaller variances.

13.2.3 Efficient estimators

With the lower bound on the variance of an estimator, we can define **efficient estimators**. An efficient estimator is optimal in the sense of using information to reduce uncertainty.

Definition 13.2.2 (efficient estimator). *An unbiased estimator $\hat{\theta}$ if variance achieves equality in the Cramer Rao lower bound for all $\theta \in \Theta$.*

The Cramer-Rao lower-bound enables us to judge whether one estimator is efficient: the closer to lower bound, the more efficient. In practice, efficient estimators are usually difficult to find. Efficient estimator is also called a **uniformly minimum-variance unbiased estimator** (UMVUE).

Example 13.2.5. Let the pmf of Bernoulli distribution parameterized by $f(x;\theta) = \theta^x(1-\theta)^{1-x}, x \in \{0,1\}$. Then

$$\begin{aligned} I(\theta) &= -E[\frac{\partial^2 \log(f(x;\theta))}{\partial\theta^2}] \\ &= E[\frac{x}{\theta^2} + \frac{1-x}{(1-\theta)^2}] \\ &= \theta(1/\theta^2) + (1-\theta)(1/(1-\theta)^2) \\ &= \frac{1}{\theta(1-\theta)} \end{aligned}$$

Consider the estimator $\hat{\theta} = \overline{X}$, then

$$E[\hat{\theta}] = E[\overline{X}] = E[\sum_{i=1}^{n} X_i/n] = \theta$$

and

$$Var[\hat{\theta}] = E[\overline{X}^2] - E[\hat{\theta}]^2 = \theta/n + \theta^2(n-1)/n - \theta^2 = \theta(1-\theta)/n$$

Therefore, the variance of the estimator is achieving the lower bound and therefore efficient.

Is UMVUE always desirable? UMVUE restricts estimator to be unbiased. However, in practice, there are many biased estimators that have smaller MSE than UMVUE.

13.2.4 Asymptotic normality and efficiency of MLE

So far, we have addressed the asymptotic unbiasedness, or consistence, of MLEs [Theorem 13.1.3], with the new tool in Fisher information, we would like to examine the asymptotic efficiency of MLEs

Theorem 13.2.6 (asymptotic normality of MLE). *[3, p. 553][4, p. 478] Let $\hat{\theta}$ be the MLE of coefficient associated with distribution $f(x;\theta)$. Let θ_0 be the true value of the parameter. Let $I_1(\theta)$ be the Fisher information matrix associated with distribution $f(x;\theta)$. It follows that MLEs are asymptotic normal; that is, in distribution,*

$$\sqrt{n}(\hat{\theta} - \theta_0) \to MN(0, [I_1(\theta_0)]^{-1}),$$

as $n \to \infty$.

It is easy see that MLE is **asymptotically efficient** because its asymptotic variance reaches the Cramer-Rao lower bound.

13.3 Sufficiency and data reduction

13.3.1 Sufficient estimators

When we estimate a model parameter θ, **not all the information in the data are relevant** to the estimation procedure. For example, if we want to estimate the mean, then the order of the sample is irrelevant. A **sufficient statistic** for a model parameter θ represents the **summary of all information from the data that are useful** for estimation of θ.

Definition 13.3.1 (sufficient statistics). *Let X be a random sample of size n. A statistic $T(X)$ is a sufficient statistic for θ if the conditional distribution of the sample X given the value of $T(X)$ does not depend on θ; that is*

$$P(X|T,\theta) = P(X|T)$$

In otherwise, X and θ are conditional independent given T.

Remark 13.3.1 (sufficient statistic as a lossless data compression). A statistic is sufficient means that $T(X)$ itself can capture all the information useful in estimating θ; the sample X might contain more information than $T(X)$(since $T(X)$ is usually not 1-1), but this additional information does not provide additional usefulness in estimating θ.

Example 13.3.1 (**trivial sufficient statistic**). The statistic $T(X_1, ..., X_n) = (X_1, ..., X_n)$ is always sufficient for any estimation task.

Example 13.3.2. Suppose $X_1, X_2 \sim B(n, \theta)$ and consider

$$P\left(X_1 = x | X_1 + X_2 = r\right)$$

$$= \frac{P\left(X_1 = x, X_1 + X_2 = r\right)}{P\left(X_1 + X_2 = r\right)}$$

$$= \frac{P\left(X_1 = x, X_2 = r - x\right)}{P\left(X_1 + X_2 = r\right)}$$

$$= \frac{\binom{n}{x}\theta^x (1-\theta)^{n-x}\binom{n}{r-x}\theta^{r-x}(1-\theta)^{n-r+x}}{\binom{2n}{r}\theta^r (1-\theta)^{2n-r}}$$

$$= \frac{\binom{n}{x}\binom{n}{r-x}}{\binom{2n}{r}}.$$

This does not contain θ, so that $X_1 + X_2$ is a sufficient statistic for θ.

13.3.2 Factorization theorem

Theorem 13.3.1 (Neyman-Fisher Factorization theorem). *[2, p. 276] Let $f(\boldsymbol{x}|\theta)$ denote the joint pdf or pmf of a sample $\boldsymbol{X}$. A statistic $T(\boldsymbol{X})$ is a sufficient statistic for θ if and only if there exist functions $g(T(\boldsymbol{x})|\theta))$ such that for all sample points $\boldsymbol{x} \in \mathcal{X}$ and all parameter points $\theta \in \Theta$, we have*

$$f(\boldsymbol{x}|\theta) = g(T(\boldsymbol{x})|\theta)h(\boldsymbol{x})$$

Proof. (1) Assume $T(\boldsymbol{X})$ is sufficient, then we have $f(\boldsymbol{x}|T(\boldsymbol{x}), \theta) = f(\boldsymbol{x}|T(\boldsymbol{x}))$. Then we have

$$\begin{aligned} f(\boldsymbol{x}|\theta) = f(\boldsymbol{x}|\theta)f(T(\boldsymbol{x})|\boldsymbol{x}, \theta) &= f(\boldsymbol{x}, T(\boldsymbol{x})|\theta) = f(T(\boldsymbol{x})|\theta)f(\boldsymbol{x}|T(\boldsymbol{x}), \theta) \\ &= f(T(\boldsymbol{x})|\theta)f(\boldsymbol{x}|T(\boldsymbol{x}))(\text{use sufficiency}) \\ &= h(\boldsymbol{x})g(T(\boldsymbol{x})|\theta) \end{aligned}$$

(2) Assume the factorization holds. Let $T(\boldsymbol{x}) = a$.

Because $f(\boldsymbol{x};\theta) = g\left(T(\boldsymbol{x})|\theta\right) h(\mathbf{x})$, we have

$$P\left(T(\boldsymbol{X}) = a\right) = \int_{\boldsymbol{y}\in T^{-1}(a)} p(\boldsymbol{y})d\boldsymbol{y} = g(a|\theta) \int_{\boldsymbol{y}\in T^{-1}(a)} h(\boldsymbol{y})d\boldsymbol{y}.$$

Hence

$$P\left(\boldsymbol{X} = \boldsymbol{x} \mid T(\boldsymbol{X}) = a\right) = \frac{h(\boldsymbol{x})}{\int_{\boldsymbol{y}\in s^{-1}(a)} h(\boldsymbol{y})d\boldsymbol{y}}$$

and this does not depend upon θ.

□

Example 13.3.3. $X_1, X_2, \ldots, X_n \sim U(0,\theta)$, so that

$$f(\boldsymbol{x}|\theta) = \theta^{-n}, \quad 0 < x_1, \ldots, x_n < \theta.$$

Or equivalently that

$$f(\boldsymbol{x}|\theta) = \theta^{-n}, \quad \theta > x_{(n)} \triangleq \max_i X_i.$$

We can factorize this as $T(\mathbf{x}) = x_{(n)}$ and $h\left(\mathbf{x}\right) = 1$, so that $X_{(n)}$ is a sufficient statistic for θ.

Example 13.3.4. Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a Bernoulli distribution. Then

$$f(\mathbf{x};\theta) = \theta^{\sum_i x_i}(1-\theta)^{n-\sum_i x_i}.$$

We can factorize this as $T(\mathbf{x}) = \sum_i x_i$ and $h\left(\mathbf{x}\right) = 1$.

Example 13.3.5. Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a $N(\mu,\sigma^2)$ distribution, where $(\mu,\sigma^2)^T$ is a vector of unknown parameters. Then

$$f(\mathbf{x};\mu,\sigma^2) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2\right]$$

$$= \left(2\pi\sigma^2\right)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_i (x_i-\overline{x})^2 + n(\overline{x}-\mu)^2\right].$$

We can factorize this as $T(\mathbf{x}) = \left(\overline{x}, \sum_i (x_i - \overline{x})^2\right)^T$.

13.4 Bayesian estimation theory

13.4.1 Overview

In our previous statistical estimation approach, we have a general setup: we are given samples drawn from a population with unknown distribution; we assume a parametric distribution and estimate the distribution parameters from the samples via properly designed estimators. In this approach, the distribution parameters θ are assumed to be some unknown and non-random quantities. This approach is generally referred to as **frequentist** approach.

In this section, we introduce an alternative approach, known as **Bayesian** approach. In the Bayesian framework, we assume the distribution parameter θ is a random variable following a prior distribution. After observing some samples, we update the distribution of θ and yield the posterior distribution of θ. The update step is usually carried out using Bayes' Rule.

Advantages of this Bayesian approach include

- It provides a natural and principled way of combining prior information with data. Such prior can incorporate expert domain knowledge or act as a form of model regularization.
- It provides uncertainty measure of the estimated parameters. For example, Bayesian approach can offer answers like the true parameter has a probability of 0.9 of falling in an interval.

The downsides of Bayesian approach include

- Additional efforts to select a prior. Mistakenly specified priors can give misleading conclusions.
- Higher computational cost, particularly for models with a large number of parameters. Bayesian approach usually require the usage of computational intensive simulation methods, i.e., Monte Carlo, to perform calculation.

13.4.2 Basics

Let $X_1, \ldots, X_n$ be random samples from a distribution. A Bayesian statistical model is composed of a **data generation model**, $X_i \sim p(x|\theta)$, and a **prior distribution model** on the model parameters, $p(\theta), \theta \in \mathbb{R}^n$. Using Bayes' theorem, the posterior distribution of θ given the data is

$$\pi\left(\theta|X_1, \ldots, X_n\right) = \frac{p\left(X_1, \ldots, X_n|\theta\right)\pi(\theta)}{m\left(X_1, \ldots, X_n\right)}$$

where

$$m\left(X_1,\ldots,X_n\right)=\int p\left(X_1,\ldots,X_n\mid\theta\right)\pi(\theta)d\theta.$$

Because m does not depend on θ, we can write

$$p\left(\theta|X_1,\ldots,X_n\right)\propto L(\theta)\pi(\theta)$$

where $L(\theta)=p\left(X_1,\ldots,X_n\mid\theta\right)$ is the likelihood function [Definition 13.1.8]. The interpretation is that $p\left(\theta|X_1,\ldots,X_n\right)$ represents the update on the subjective beliefs about θ after observing $X_1,\ldots,X_n$.

With posterior distribution, two commonly used estimators are

- Posterior mean estimator

$$\bar{\theta}=\mathbb{E}\left(\theta\mid X_1,\ldots,X_n\right)=\int\theta\pi\left(\theta\mid X_1,\ldots,X_n\right)d\theta=\frac{\int\theta L(\theta)\pi(\theta)}{\int L(\theta)\pi(\theta)}.$$

- Maximum A posterior estimator is given as

$$\hat{\theta}_{MAP}(x)=\arg\max_{\theta}p(\theta|x)=\arg\max_{\theta}p(x|\theta)p(\theta).$$

Example 13.4.1. Let the data generation model be $Bernoulli(\theta)$ and the prior distribution on p be $\theta\sim Beta(\alpha,\beta)$. Let $X=X_1,\ldots,X_n$ be random samples. Then

$$\pi(\theta)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)},\Gamma(\alpha)=\int_0^\infty t^{\alpha-1}e^{-t}dt.$$

Set $Y=\sum_i X_i$. Then

$$\pi(p|X)\propto\underbrace{\theta^Y(1-\theta)^{n-Y}}_{\text{likelihood}}\times\underbrace{\theta^{\alpha-1}(1-\theta)^{\beta-1}}_{\text{prior}}\propto\theta^{Y+\alpha-1}(1-\theta)^{n-Y+\beta-1}.$$

Therefore, $p|X\sim Beta(Y+\alpha,n-Y+\beta)$. The Posterior mean estimator [Lemma 12.1.28] is

$$\hat{p}=\frac{Y+\alpha}{(Y+\alpha)+(n-Y+\beta)}=\frac{Y+\alpha}{\alpha+\beta+n}.$$

In the above example, the prior and posterior distribution belong to the same family. This is an example of a conjugate prior.

Definition 13.4.1 (conjugate prior). *$p(\theta)$ is a conjugate prior for $p(x|\theta)$ if*

$$p(\theta) \in \mathcal{P} \Rightarrow p(\theta|x) \in \mathcal{P}$$

where $\mathcal{P}$ is a family of pdf parameterized by θ. In other words, $p(\theta)$ and $p(x|\theta)$ are in the same family.

It is beneficial to use use conjugate prior. When prior and posterior distribution are in the same family, it is easy to interpret how the observations x changes the prior distribution. For a comprehensive account of conjugate priors, see [5][6].

Example 13.4.2. Let $X_1, \ldots, X_n \sim N\left(\mu, \sigma^2\right)$ with σ^2 known. Let $\mu \sim N\left(m, \tau^2\right)$. Then the posterior distribution is given by

$$p(\mu|X) \propto \exp(\frac{\sum_{i=1}^{N}(\mu - X_i)^2}{2\sigma^2}) \exp(\frac{(\mu - m)^2}{2\tau^2}).$$

The maximum a posterior estimator $\hat{\mu}$ is given by maximizing posterior estimator

$$\hat{\mu} = \frac{\frac{\sum_{i=1}^{N} X_i}{\sigma^2} + \frac{m}{\tau^2}}{\frac{N}{\sigma^2} + \frac{1}{\tau^2}}.$$

13.5 Bootstrap method

Let $\boldsymbol{X} = (X_1, X_2, ..., X_n)$ be a random sample of size n and let $\hat{\theta}(\boldsymbol{X})$ be a statistic of interest. One central task of statistical estimation is characterize the variance of $\hat{\theta}(\boldsymbol{X})$. In simple cases, we might be able to directly derive the distribution of the estimator. For example, let $\boldsymbol{X}$ be random samples of a normal distribution, the sample variance S^2 will have $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$.In complex cases where obtaining standard deviation, confidence interval or even distributions of $\hat{\theta}$ is difficult. The goal of bootstrap methods is to measure the standard deviation, confidence interval, or even distributions of $\hat{\theta}$ by numerical simulation method.

On a high level, a bootstrap method of estimating the variance of an estimator consisting of the following steps

- Draw B bootstrap samples, a bootstrap sample is a set of N sample drawn from the original samples with replacement.
- On each bootstrap sample i, evaluate the estimator $\hat{\theta}(\boldsymbol{X})_i$.
- Estimate the variance of $\hat{\theta}(\boldsymbol{X})$ from $\hat{\theta}(\boldsymbol{X})_i, i = 1, ..., B$.

The intuition of the working mechanism underlying bootstrap method is the fact that we can view the bootstrap sample as a new set of samples drawn from the empirical sample distribution (the joint distribution of $(X_1, ..., X_N)$).

Remark 13.5.1 (resampling property). Given a sample of size n, we re-draw n sample with replacement. Then from probability, we have:

- The probability that ith sample is not being resampled is $(1 - \frac{1}{n})$ at the first time.
- The probability that ith sample is not being resampled is $(1 - \frac{1}{n})^n$ at the new sample of size n.
- The probability that ith sample is not being resampled is e^{-1} at the new sample when $n \to \infty$.
- On average, about ne^{-1} of original samples will not show in the new sample as $n \to \infty$.

Now we can summarize the basic procedure in a bootstrap method.

Methodology 13.5.1 (general bootstrap estimation). *Let $\hat{\theta}$ be a statistic as a function of $(X_1, ..., X_N)$. Let $\hat{\theta}_i$ be the estimation evaluated at bootstrap sample i. Then*

- *The mean estimation*

$$m = \frac{1}{B} \sum_{i=1}^{B} \hat{\theta}_i \approx E[\hat{\theta}].$$

- *The variance estimation is*

$$s = \frac{1}{B}\sum_{i=1}^{B}(\hat{\theta}_i - m)^2 \approx Var[\hat{\theta}].$$

Clearly, there are two sources of error in the bootstrap estimate: The first arises from finite sample size N, and the second arises from finite B. In practice, we usually take as large as possible, say B = 10000 or $\sim N^2$. As $B \to \infty$, $Var[\hat{\theta}]$ will converge.

By properly modifying the variance estimation procedure, we arrive at the following confidence level estimation method.

Methodology 13.5.2 (bootstrap confidence level). *Let $\hat{\theta}$ be a statistic as a function of $(X_1, ..., X_N)$. Let $\hat{\theta}_i$ be the estimation evaluated at bootstrap sample i. Let $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_B$ be sorted. Denote $k_1 = (B \times \frac{\alpha}{2}), k_2 = (B \times (1 - \frac{\alpha}{2}))$. Then $[\hat{\theta}_{k_1}, \hat{\theta}_{k_2}]$ is the α confidence interval such that*

$$Pr(\hat{\theta}_{k_1} \leq \theta \leq \hat{\theta}_{k_2}) = 1 - \alpha.$$

Remark 13.5.2. One variation of bootstrap method is the jackknife where the standard error is estimated by from $N - 1$ leaving-one-out subsamples.

13.6 Hypothesis testing theory

13.6.1 Basics

In statistical modeling of observed data, one may put forward a hypothesis regarding the specification of statistical models, statistical relationship among data or between different group of data, etc.

For example, a principal of a school claims that the students in his school have an average height at 5 feet. Suppose measurement of heights of 100 students, we get average height of 5.5 feet and standard deviation of 0.5 feet. Is there sufficient evidence to conclude the principal's statement?

In a typical hypothesis testing, we usually propose two **hypotheses**: **null hypothesis** and **alternative hypothesis**, denoted as H_0 and H_1. In general, null hypothesis and the alternative hypothesis are **complementary** to each other. Our goal is to determine which one is statistically correct or incorrect.

The null hypothesis is usually a simple hypothesis **the contradiction** to what we would like to prove.The alternative hypothesis is usually a hypothesis what we would like to prove. Alternative hypothesis can be **two-sided or one-sided**.

Example 13.6.1. Consider a clinical trial of a new drug. Treatment data are collected to compare two treatments.
The *null hypothesis* is usually no difference between treatments.

Depending on the purpose the test, the *alternative hypothesis* might be:

- New drug is different from old drug. (*two-sided*)
- New drug is better than old drug.(*one-sided*),
- Old drug is better than new drug. (*two-sided*).

Example 13.6.2. Given observation sampled from a normal distribution. A null hypothesis regarding the mean μ can be $\mu = \mu_0$.

And the alternative hypothesis can be

- $H_1 : \mu > \mu_0$, which is an **upper-tailed one-sided hypothesis**.
- $H_1 : \mu < \mu_0$, which is a **lower-tailed one-sided hypothesis**.
- $H_1 : \mu \neq \mu_0$, which is a **two-sided hypothesis**.

Mathematically, a hypothesis can be viewed as a statement about a population parameter θ. Let θ denote a population parameter. The general format of the null and

alternative hypothesis is $H_0 : \theta \in \Theta_0$, and $H_1 : \theta \in \Theta_1$, where Θ_0 and Θ_1 are disjoint subsets of the parameter space Θ.

Example 13.6.3. A coin is tossed and we hypothesize that it is fair. Hence Θ_0 is the set $\left\{\frac{1}{2}\right\}$ containing just one element of the parameter space $\Theta = [0, 1]$

Given H_0 and H_1, we need to decide which hypothesis to reject or accept, or which partition in Θ the population coefficient θ lies in. From this perspective, we can view hypothesis testing as a decision making problem with uncertainty.

Usually, decision is based on $p(\theta|x)$, the posterior distribution can be calculation as $p(x|\theta \in H_i)$. Given the observation data we can calculate $p(x|H_0)$ and $p(x|H_1)$. We will determine which, H_0 or H_1, is more appropriate, by either comparing $p(x|H_0)$ and $p(x|H_1)$ or specifying the value range of a test statistics T to accept or reject H_0.

A basic framework of hypothesis testing using test statistic can be summarized in the following method.

Methodology 13.6.1 (hypothesis testing via test statistic). *Suppose we are given random samples drawn from a normal distribution with known various σ^2 and unknow mean μ. A typical hypothesis testing on the μ involves the following procedures.*

- *State the Null hypothesis. For example $H0 : \mu = \mu_0$.*
- *State the Alternate Hypothesis. For example, $H1 : \mu > \mu_0$ or $\mu < \mu_0$ or $\mu \neq \mu_0$.*
- *State the significance level α, say $\alpha = 0.05$.*
- *Select the appropriate test statistic. For example,*

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}.$$

- *Determine the rejection region area [Figure 13.6.1]. For test statistic Z and $H_1 : \mu > \mu_0$, one rejection region could be $R = \{Z : Z > \phi^{-1}(1-\alpha)\}$, where ϕ^{-1} is the inverse cdf of Z. Rejection regions for other cases can be determined similarly.*
- *Calculate the test statistic value and accept or reject H_0 based on Z. If $Z \in R$, we reject the null hypothesis.*

The **significance level** α is a probability threshold below which the **null hypothesis will be rejected under the assumption that H_0 is true**. Common values are 0.05 and 0.01.

There are different ways to specify a decision rule. In general, the decision rule depends on whether the test is an upper-tailed, lower-tailed, or two-tailed test[Figure 13.6.1]. In an upper-tailed or lower-tailed test the rejection region is around the upper or lower

tail whereH_0 will will get rejected when the test statistic is larger or smaller than a critical value, respectively. In a two-tailed test, the rejection region is at both tails where H_0 will get rejected if the test statistic is extreme, either larger than an upper critical value or smaller than a lower critical value.

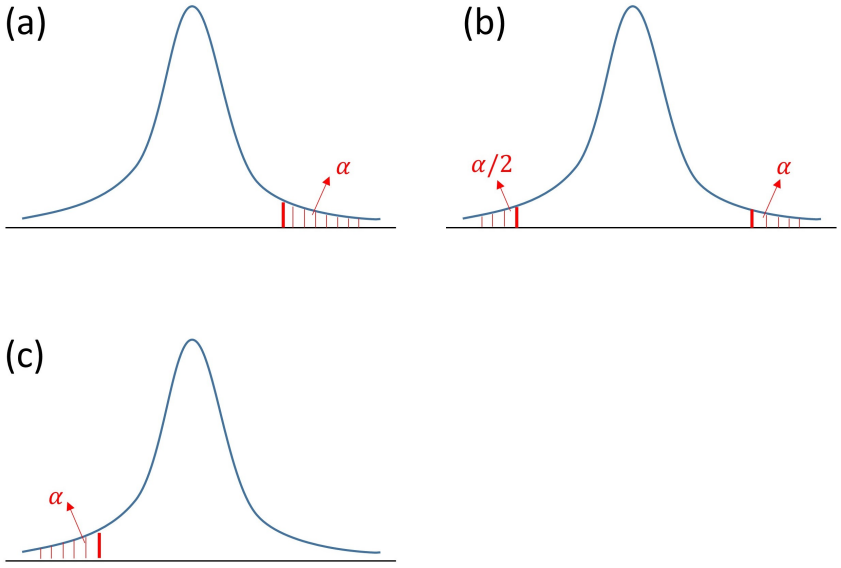


Figure 13.6.1: Demonstration for rejection regions for upper-tailed one-sided hypothesis (a), two-sided hypothesis(b), and lower-tailed one-sided hypothesis (c).

Definition 13.6.1 (p value). *The p value is the probability, assuming the null hypothesis is true, of observing the at least as extreme as (equal to or "more extreme" than) the observed test statistic in the alternative hypothesis direction.*
p value can also be interpreted as the smallest significant value that H_0 will be reject.

Methodology 13.6.2 (p value method). *Given a significance level α:*

- *If $p \leq \alpha$, then reject H_0.*
- *If $p > \alpha$, then accept H_0.*

Example 13.6.4. Consider a hypothesis test of n random samples from normal distribution $N(\mu, \sigma^2)$. Let $H_0 : \mu = \mu_0$. Let the test statistic be

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}.$$

- If $H_1 : \mu > 0$, the $p = 1 - \Phi^{-1}(Z)$. If $p \leq \alpha$, we reject H_0.
- If $H_1 : \mu < 0$, the $p = 1 - \Phi^{-1}(-Z)$. If $p \leq \alpha$, we reject H_0.
- If $H_1 : \mu \neq 0$, the $p = 2(1 - \Phi^{-1}(|Z|))$. If $p \leq \alpha$, we reject H_0.

where Φ is the cdf of a standard normal distribution.

13.6.2 Characterizing errors and power

We usually only test if H_0 is true or not and **do not test** the correctness H_1. For a binary hypothesis testing, there could be four types of results.

Definition 13.6.2 (Four types of results in binary hopothesis testing).

1. *Detection: H_0 true, decide H_0*
2. *False alarm/ **type I error:** H_0 true, we reject H_0, decide H_1.*
3. *Miss/ **type II error:** H_1 true, decide H_0(or H_0 is false, we do not reject H_0.)*
4. *Correctly rejection: H_1 true,decide H_1*

There are two types of possible error.

- A **Type I error** is the error of rejecting the null hypothesis H_0 when H_0 is true.
- A **Type II error** is the error of not rejecting the null hypothesis H_0 when H_0 is false.

We have following summary table.

	H_0 not rejected	H_0 rejected
H_0 true	no error	Type I error
H_0 false	Type II error	no error

We usually denote

$$\alpha = Pr(\text{Type I error}), \beta = Pr(\text{Type II error}).$$

Example 13.6.5 (type I, II error in hypothesis test of a normal distribution). Consider a hypothesis test of n random samples from normal distribution $N(0, \sigma^2)$. Let $H_0 : \mu = 0, H_1 : \mu > 0$. Let the test statistic be

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}.$$

Then

- To calculate type I error, we assume $H_0 : \mu = 0$ is correct, then

$$Pr(\text{type I error}) = Pr(\frac{\bar{X}}{\sigma/\sqrt{n}} > z_\alpha | H_0) = \alpha,$$

  where z_α is defined as $\Phi(z_\alpha) = 1 - \alpha$, $\Phi(z)$ is the cdf of a normal distribution.
- To calculate type II error, we assume $H_1 : \mu > 0$ is correct, then $Pr(\text{type II error}) = Pr(Z < z_\alpha | H_1)$ can be calculated in the following way:

$$\begin{aligned}\beta &= P(\frac{\bar{X}}{\sigma/\sqrt{n}} < z_\alpha) \\ &= P(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_\alpha - \frac{\mu}{\sigma/\sqrt{n}}) \\ &= \Phi(z_\alpha - \frac{\mu}{\sigma/\sqrt{n}})\end{aligned}$$

  There are several critical implications:
  - If μ increases, then β decreases.
  - If n increases, then β decreases.
  - If α increases, then z_α increases and β increases.

Remark 13.6.1 (interpretation on two types of errors).

- Under the null hypothesis H_0, significance level α is the probability measure (i.e., the size) of the rejection region where H_0 will be rejected. α bounds the type I error.
- Given a fixed size of samples, it is generally not possible to minimize both types of error.
- We usually consider type I error to be worse and try to minimize or bound type I error first and then minimize type II error.
- H_0 is usually conservative statement such that reject H_0 when it is true will have **significant bad consequence**.

13.6.3 Power of a statistical test

Hypothesis testing inherently involves two type of test errors, which usually can not be minimized together. In practice, we design hypothesis and choosing signficance level following the principle that

- **Minimize the probability of committing a Type I error.** That, is minimize $\alpha = P(TypeIError)$. Typically, $\alpha \leq 0.1$.
- **Maximize the power, or reduce the type II error** Note that $\beta = P(TypeIIError) = 1 - power$, typically $\beta \leq 0.2$.

In this section, we will give a close look at the statistical power.

Definition 13.6.3 (statistical power of a test). *The power of a hypothesis test is the probability of making the correct decision if the alternative hypothesis is true. That is, the power of a hypothesis test is the probability of rejecting the null hypothesis H_0 is incorrect(or when the alternative hypothesis H_1 is true):*

$$power = P(reject\ H_0|H_1).$$

Let's revisit the previous example. Consider a hypothesis test of n random samples from normal distribution $N(0, \sigma^2)$. Let $H_0 : \mu = 0, H_1 : \mu > 0$. Let the test statistic be

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

To calculate type II error, we assume $H_1 : \mu > 0$ is correct, then $Pr(\text{type II error}) = Pr(Z < z_\alpha|H_1)$ can be calculated in the following way:

$$\begin{aligned}\beta &= P(\frac{\bar{X}}{\sigma/\sqrt{n}} < z_\alpha) \\ &= P(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_\alpha - \frac{\mu}{\sigma/\sqrt{n}}) \\ &= \Phi(z_\alpha - \frac{\mu}{\sigma/\sqrt{n}})\end{aligned}$$

Since power equals $1 - \beta$, there are several critical implications:

- If μ increases, then β decreases, and power increases
- If n increases, then β decreases, and power increases
- If α increases, then z_α increases, β increases, power decreases

We have the following summary on factors affecting statistical power.

Note 13.6.1 (factors affecting statistical power). Statistical power may depend on a number of factors.

- the statistical significance criterion used in the test, i.e, α.
- the magnitude of the effect of interest in the population, i.e., μ.
- the sample size used to detect the effect, i.e., n.

Power analysis can be used to calculate the minimum sample size required so that one can be reasonably likely to detect an effect of a given size. For example: "how many times do I need to toss a coin to conclude it is unfair?

Example 13.6.6 (calculating required sample size). Following previous example, the power in a normal distribution mean test is given by

$$power = 1 - \Phi(z_\alpha - \frac{\mu}{\sigma/\sqrt{n}}),$$

If we need power to be greater than p_0, then via algebra, we can get

$$n \geq \frac{\sigma^2}{\mu^2}(z_\alpha - \Phi^{-1}(1-p_0)).$$

13.6.4 Common statistical tests

13.6.4.1 *Chi-square goodness-of-fit test*

Theorem 13.6.1 (Pearson's theorem). *Consider r boxes $B_1, ..., B_r$ and throw n balls $X_1, X_2, ..., X_n$ into these boxes independently of each other with probabilities*

$$P(X_1 \in B_1) = p_1, ..., P(X_r \in B_r) = p_r,$$

such that $p_1 + ... + p_r = 1$.

Let v_j be the number of balls in the jth box, i.e. $v_j = \sum_{i=1}^n \mathbf{1}_{X_i = B_j}$.

It follows that

- *The random variable*

$$\frac{v_j - np_j}{\sqrt{np_j}} \to N(0, 1-p_j) \text{ in distribution, as } n \to \infty$$

- *The random vector $Y = (Y_1, Y_2, ..., Y_r)$, $Y_j = \frac{v_j - np_j}{\sqrt{np_j}}$ will converge to $MN(0, \Sigma)$ in distribution, where*

$$\Sigma_{ii} = 1 - p_i, \Sigma_{ij} = -\sqrt{p_i p_j}.$$

- *The random variable*

$$\sum_{j=1}^r \frac{(v_j - np_j)^2}{np_j} \to \chi^2(r-1) \text{ in distribution, as } n \to \infty.$$

Proof. (1) Note that from Bernoulli distribution

$$E[\mathbf{1}(X_1 \in B_j)] = p_j, Var[\mathbf{1}(X_1 \in B_j)] = p_j(1-p_j).$$

By the central limit theorem

$$\frac{v_j - np_j}{\sqrt{np_j(1-p_j)}} \to N(0,1) \text{ in dist } \implies \frac{v_j - np_j}{\sqrt{np_j}} \to N(0,1-p_j) \text{ in dist.}$$

(2)

$$\begin{aligned} E[\frac{v_i - np_i}{\sqrt{np_i}} \frac{v_j - np_j}{\sqrt{np_j}}] &= \frac{1}{n\sqrt{p_i p_j}}(E[v_i v_j] - n^2 p_i p_j) \\ E[v_i v_j] &= E[\sum_{l=1}^{n} \mathbf{1}(X_l \in B_i) \sum_{k=1}^{n} \mathbf{1}(X_k \in B_i)] \\ &= E[\sum_{l=1}^{n} \sum_{k=1, k\neq l}^{n} \mathbf{1}(X_l \in B_i)\mathbf{1}(X_k \in B_i)] \\ &= 2E[\sum_{l=1}^{n} \sum_{k>1}^{n} \mathbf{1}(X_l \in B_i)\mathbf{1}(X_k \in B_i)] \\ &= n(n-1)p_i p_j \\ E[\frac{v_i - np_i}{\sqrt{np_i}} \frac{v_j - np_j}{\sqrt{np_j}}] &= -\sqrt{p_i p_j}. \end{aligned}$$

(3) Note that

$$Y^T Y = Z^T(I - UU^T)Z, Z \in MN(0, I_r), U = (\sqrt{p_1}, \sqrt{p_2}, ..., \sqrt{p_r}),$$

where UU^T is an rank 1 orthogonal projector($U^T U = p_1 + p_2 + \cdots + p_r = 1$).

From the chi-square decomposition theorem [Theorem 12.4.1], we know that $Y^T Y \to \chi^2(r-1).in.dist.$ □

Theorem 13.6.2 (chi-square goodness-of-fit test). *Suppose that we observe an iid sample $X_1, X_2, ..., X_n$ of random variable that take a finite number of values $B_1, B_2, ..., B_r$ with unknown probabilities $p_1, p_2, ..., p_r$. Consider hypotheses*

$$H_0 : p_i = p_i^0, for\ i = 1, 2, ..., r$$
$$H_1 : for\ some\ i, p_i \neq p_i^0$$

and the test statistic

$$T = \sum_{i=1}^{r} \frac{(v_i - np_i^0)^2}{np_i^0},$$

where $v_j = \sum_{i=1}^{n} \mathbf{1}_{X_i = B_j}$.

It follows that

- *If H_0 is true, then $T \to \chi^2(r) - 1$ in dist.*
- *If H_1 is true, then $T \to \infty$, as $n \to \infty$.*
- *The decision rule is reject H_0 if $T > c$ where $c = \inf\{z : F(z) \geq 0.99\}$.*

Proof. (1) From Pearson's theorem [Theorem 13.6.1]. (2) If we write

$$\frac{(v_i - np_i^0)}{\sqrt{np_i^0}} = \sqrt{\frac{p_i}{p_i^0}} \frac{(v_i - np_i)}{\sqrt{np_i^0}} + \sqrt{n} \frac{(v_i - n(p_i - p_i^0)}{\sqrt{np_i^0}},$$

then the second quantity will diverge as $n \to \infty$. □

Note 13.6.2 (p value method for chi-square test)**.** The p-value for a chi-square test is defined as the **tail area above the calculated test statistic**.

For example, consider an experiment with test statistic result

$$T = \sum_{i=1}^{r} \frac{(v_i - np_i^0)^2}{np_i^0}.$$

Then

$$p - value = Pr(\chi^2(r-1) \geq T).$$

Given a significance level α:

- If $p \leq \alpha$, then reject H_0.
- If $p > \alpha$, then accept H_0.

13.6.4.2 *Chi-square test for statistical independence*

Lemma 13.6.1. *link*

Denote

$$p_i = \sum_{j=1}^{c} \frac{O_{ij}}{N}, q_j = \sum_{i=1}^{r} \frac{O_{ij}}{N}, E_{ij} = N p_i q_j$$

$$\sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{i,j} - E_{ij})^2}{E_{ij}} \sim \chi^2(p),$$

where $p = (r-1)(c-1)$.

The hypothesis is given by

- *H_0: U is independent of V;*

- *H_1: there exists an statistical relationship between U and V.*

13.6.4.3 *Kolmogorov-Smirnov goodness-of-fit test*

Definition 13.6.4 (Kolmogorov-Smirnov(KS) goodness-of-fit test). *The Kolmogorov-Smirnov goodness-of-fit test for a random sample of size N has the following elements:*

- *Hypothesis:*
  - *H_0: the data follow a specified **continuous** distribution with cdf $F(t)$.*
  - *H_1: the data do not follow the specified distribution.*
- *For **ascending ordered** sample $Y_1, Y_2, ..., Y_N$. KS test statistic is defined as*

$$D = \max_{1 \le i \le N} (F(Y_i) - \frac{i-1}{N}, \frac{i}{N} - F(Y_i)).$$

- *The significance level α and critical value K_α.*
- *If $D > K_\alpha$, reject H_0.*

Remark 13.6.2 (interpretation and usage).

- The KS test statistic is measuring the distance of proposed distribution F is the empirical cdf given by $(i-1)/N$ and i/N.
- KS test is used for continuous distribution test. For discrete distribution test, see chi-square goodness-of-fit test [Theorem 13.6.2].
- For the KS critical value table, see link.

13.7 Hypothesis testing on normal distributions

Common notations in this sections:

- sample mean $\overline{X}$
- sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - E[X])^2$

13.7.1 Normality test

Before we proceed to hypothesis testing related to normal distributions, we need to review typical methods used to determine if samples are drawn from a normal distribution. Most straight forwards methods are qualitative graphical methods in where we plot the histogram or QQ plots. In QQ plot, we plot the quantiles of the sample against the theoretical quantiles of a standard normal distribution. For sample truly drawn from a normal distribution, we expect the plot is a perfect straight line; Different deviation from a straight line will be observed when the sample distribution is has non-zero excess Kurtosis (heavy tails or not) or skewness [Figure 13.7.1].

Additional quantitative testing methods include Shapiro–Wilk test, Kolmogorov–Smirnov test, Jarque–Bera test, Pearson's chi-squared test Theorem 13.6.1, D'Agostino's K-squared test, etc.

13.7.2 Sample mean with known variance

Consider we have n samples $X_1, ..., X_n$ for a random variable with $N(\mu, \sigma^2)$ with σ^2 known. The hypothesis testing involving the mean can be obtained by using the fact that

$$Z = \frac{\overline{X} - \mu}{\sigma / n}$$

is $N(0, 1)$. We can summarize the test as:

Table 13.7.1: Test on mean with known variance σ^2

H_0	test statistic	H_1	critical(rejection) region
$\mu \leq \mu_0$	$Z = \frac{\overline{X}-\mu}{\sigma/n}$	$\mu > \mu_0$	$z \geq z_\alpha$
$\mu \geq \mu_0$	$Z = \frac{\overline{X}-\mu}{\sigma/n}$	$\mu < \mu_0$	$z \leq -z_\alpha$
$\mu = \mu_0$	$Z = \frac{\overline{X}-\mu}{\sigma/n}$	$\mu \neq \mu_0$	$\lvert z \rvert \geq z_\alpha$

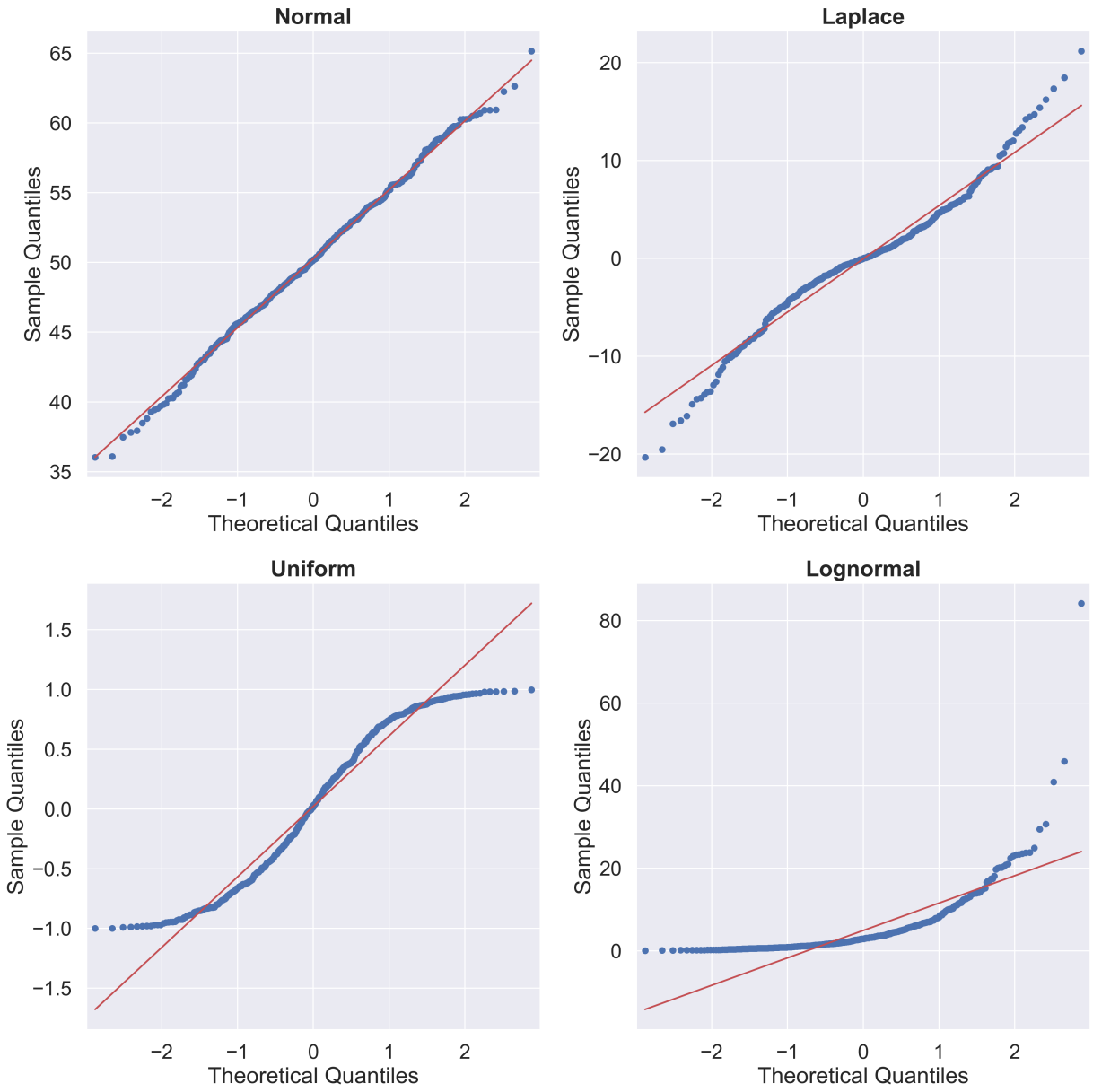


Figure 13.7.1: QQ plot with different sample distributions, including normal, Laplace ($b = 4$), Uniform $U([-1,1])$, Lognormal $LN(0,1)$. Red solid lines are the fitted linear lines.

13.7.3 Sample mean with unknown variance

Consider we have n samples $X_1, ..., X_n$ for a random variable with $N(\mu, \sigma^2)$ with σ^2 being unknown. The hypothesis testing involving the mean can be obtained by using the fact that

$$T = \frac{\overline{X} - \mu}{\sigma / n}$$

is $t(n-1)$. [Theorem 12.4.2] We can summarize the test as:

Table 13.7.2: Test on mean with unknown variance σ^2

H_0	test statistic	H_1	critical(rejection) region
$\mu \leq \mu_0$	$T = \frac{\overline{X}-\mu}{S/n}$	$\mu > \mu_0$	$t \geq t_\alpha(n-1)$
$\mu \geq \mu_0$	$T = \frac{\overline{X}-\mu}{S/n}$	$\mu < \mu_0$	$t \leq -t_\alpha(n-1)$
$\mu = \mu_0$	$T = \frac{\overline{X}-\mu}{S/n}$	$\mu \neq \mu_0$	$\lvert t \rvert \geq t_\alpha(n-1)$

13.7.4 Variance test

Consider we have n samples $X_1, ..., X_n$ for a random variable with $N(\mu, \sigma^2)$ with σ^2 being unknown. The hypothesis testing involving the mean can be obtained by using the fact that

$$T = \frac{\overline{X} - \mu}{\sigma / n}$$

is $t(n-1)$.[Theorem 12.4.2] We can summarize the test as:

Table 13.7.3: Test on variance

H_0	test statistic	H_1	critical(rejection) region
$\sigma^2 \leq \sigma_0^2$	$T = \frac{(n-1)S^2}{\sigma_0^2}$	$\sigma^2 \leq \sigma_0^2$	$t \geq \chi^2_\alpha$
$\sigma^2 \geq \sigma_0^2$	$T = \frac{(n-1)S^2}{\sigma_0^2}$	$\sigma^2 \leq \sigma_0^2$	$t \leq \chi^2_{1-\alpha}$
$\sigma^2 = \sigma_0^2$	$T = \frac{(n-1)S^2}{\sigma_0^2}$	$\sigma^2 \neq \sigma_0^2$	$t \leq \chi^2_{1-\alpha}$ or $t \geq \chi^2_\alpha$

13.7.5 Variance comparison test

Consider we have n samples $X_1, ..., X_n$ for a random variable with $N(\mu, \sigma^2)$ with σ^2 being unknown. The hypothesis testing involving the mean can be obtained by using the fact that

$$T = \frac{\overline{X} - \mu}{\sigma / n}$$

is $t(n-1)$. [Theorem 12.4.2] We can summarize the test as:

Table 13.7.4: Test on variance comparison between two samples

H_0	test statistic	H_1	critical(rejection) region
$\sigma_1^2 \leq \sigma_2^2$	$F = \frac{S_1^2}{S_2^2}$	$\sigma_1^2 \leq \sigma_2^2$	$t \geq \chi_\alpha^2$
$\sigma^2 \geq \sigma_0^2$	$F = \frac{S_1^2}{S_2^2}$	$\sigma_1^2 \leq \sigma_2^2$	$t \leq \chi_{1-\alpha}^2$
$\sigma_1^2 = \sigma_2^2$	$F = \frac{S_1^2}{S_2^2}$	$\sigma_1^2 \neq \sigma_2^2$	$F \leq \chi_{1-\alpha}^2$ or $F \geq \chi_\alpha^2$

13.7.6 Person correlation t test

Lemma 13.7.1 (Person correlation t test). *Let X and Y be random variable related by $Y = \beta X + \epsilon$, where $\beta \in \mathbb{R}$ and $\epsilon \sim N(\mu, \sigma)$. Let $\hat{\rho}$ be the correlation estimated from n samples of X and Y. Let the statistic*

$$T = \frac{\hat{\rho}}{\sqrt{1-\hat{\rho}^2}}\sqrt{n-2}$$

follows t distribution with degree of freedom $n-2$.

Proof. Note that if we construct the linear regression model on $Y \sim \beta X$, then [Theorem 15.1.11]

$$\hat{\rho}^2 = \frac{\hat{\beta}^2 S_{XX}}{S_{YY}}, 1 - \hat{\rho}^2 = \frac{SSE}{S_{YY}}.$$

Therefore

$$\frac{\hat{\rho}}{\sqrt{1-\hat{\rho}^2}}\sqrt{n-2} = \sqrt{n-2}\sqrt{\frac{S_{XX}}{SSE}}\hat{\beta}$$

is the t statistics follows $n-2$ degrees of freedom [Methodology 15.1.2]. □

13.7.7 Two sample tests

13.7.7.1 *Two-sample z test*

Bbasic setup of two-sample z test:

- $X_1, X_2, ..., X_m$ is a random sample from a distribution with mean μ_1 and variance σ_1^2.
- $Y_1, Y_2, ..., Y_n$ is a random sample from a distribution with mean μ_1 and variance σ_1^2.
- X and Y samples are independent of each other.

Lemma 13.7.2 (mean difference estimator). *[7, p. 363]Let $\overline{X}$ and $\overline{Y}$ denote the sample mean.*

- *$E[\overline{X} - \overline{Y}] = \mu_1 - \mu_2$, i.e., $\overline{X} - \overline{Y}$ is the unbiased estimator of $\mu_1 - \mu_2$.*
- $$Var[(\overline{X} - \overline{Y})] = \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}.$$

Proof. (1) Straight forwward. (2) Using independence, we have

$$Var[\overline{X} - \overline{Y}] = Var[\overline{X}] + Var[\overline{Y}] = \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}.$$

□

13.7.7.2 *Two-sample t test*

Basic setup two-sample t test:

- $X_1, X_2, ..., X_m$ is a random sample from a distribution with mean μ_1 and variance σ_1^2.
- $Y_1, Y_2, ..., Y_n$ is a random sample from a distribution with mean μ_2 and variance σ_2^2.
- X and Y samples are independent of each other.

Lemma 13.7.3 (mean difference estimator). *[7, p. 363] The standardized variable*

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}}$$

has approximately a t distribution with degree of freedom v estimated to be (round to the nearest integer)

$$v = \frac{\left(\frac{S_1^2}{m} + \frac{S_2^2}{n}\right)^2}{\frac{\left(S_1^2/m\right)^2}{m-1} + \frac{\left(S_2^2/n\right)^2}{n-1}}$$

13.7.7.3 *Paired Data*

Basic setup for paired data

- The data consists of n independently selected pairs $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$, with $E[X_i] = \mu_1$ and $E[Y_i] = \mu_2$.
- Let $D_1 = X_1 - Y_1, D_2 = X_2 - Y_2, ..., D_n = X_n - Y_n$ so that $D_i's$ are the differences within pairs.
- The $D_i's$ are assumed to be normally distributed within mean value μ_D and variance σ_D^2.

Because we assume D_i are IID samples from normal distribution, we can apply the two-sample z test or two sample t test to test if D_i has zero mean.

Note 13.7.1 (caution!). X_i could be dependent on Y_i, but pairs are independent of each other. Here we assume D_i follows normal distribution, this is not necessarily true even if X and Y are normal distributions [Corollary 12.1.1.1].

13.7.8 Interval estimation for normal distribution

Definition 13.7.1 (confidence interval). *Let $X_1, X_2, ..X_n$ denote a random sample on a random variable X, where X has pdf $f(x;\theta)$. Let $\alpha (0 < \alpha < 1)$ be given. Let $L = L(X_1, X_2, ..X_n), U = U(X_1, X_2, ..X_n)$ be two statistics. We say that the interval (L, U) is a $(1-\alpha)$ confidence interval for θ if*

$$1 - \alpha = P_\theta(\theta \in (L, U))$$

Lemma 13.7.4 (confidence interval for mean of normal random sample). *Let X be a normal random variable $N(\mu, \sigma^2)$, Let $X_1, ..., X_n$ be the random sample, let S^2 and $\overline{X}$ be the sample variance and sample mean, then*

- *If σ is known, then the $(1-\alpha)$ confidence interval for μ is*

$$(\overline{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \overline{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2})$$

- *If σ is unknown, then the $(1-\alpha)$ confidence interval for μ is*

$$(\overline{X} - \frac{S}{\sqrt{n}} t_{\alpha/2}(n-1), \overline{X} + \frac{S}{\sqrt{n}} t_{\alpha/2}(n-1))$$

where $z_{\alpha/2}, t_{\alpha/2)}(n-1)$ are the upper critical point of $\alpha/2$ for standard normal distribution and $t(n-1)$ distribution.

Proof. (1) Use the fact that

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

(2) Use the fact that

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

□

Remark 13.7.1 (knowing σ reduce uncertainty). Note that t distribution is wider(has big tails) than normal, which suggest larger confidence interval when σ is unknown.

Lemma 13.7.5 (Large sample confidence interval). *[8, p. 220] Let $X_1, ..., X_n$ be the random sample of a random variable with mean μ and variance σ^2. (Note that X is not necessarily normal). Then the $(1-\alpha)$ confidence interval for μ for large sample size is given as*

$$(\overline{X} - \frac{S}{\sqrt{n}} z_{\alpha/2}, \overline{X} + \frac{S}{\sqrt{n}} z_{\alpha/2})$$

Proof. When n is large, $S \approx \sigma$. Based on central limit theorem Theorem 11.11.3.

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

□

13.8 Notes on bibliography

For an advanced treatment on statistical estimation theory, see [2][8]. For likelihood based methods, see [9]. For large sample theory(asymptotic analysis), see [10].

For introductory level Bayesian statistics, see [11].

For a good treatment on statistical estimation theory, see [12].

For a tutorial on Fisher Information matrix, see [13]

For an introduction to robust statistics, see [14].

For an extensive discussion on statistical distribution, see [15][16].

BIBLIOGRAPHY

1. Ruppert, D. *Statistics and data analysis for financial engineering, 2ed* (Springer, 2015).
2. Casella, G. & Berger, R. L. *Statistical inference* (Duxbury Pacific Grove, CA, 2002).
3. Moon, T. K. S. & Wynn, C. *Mathematical methods and algorithms for signal processing* **621.39: 51 MON** (2000).
4. Greene, W. *Econometric Analysis* ISBN: 9780134461366 (Pearson, 2017).
5. Fink, D. A compendium of conjugate priors. *See http://www. people. cornell. edu/pages/df36/CONJINTRnew% 20TEX. pdf*, 46 (1997).
6. Wikipedia. *Conjugate prior — Wikipedia, The Free Encyclopedia* [Online; accessed 7-September-2016]. 2016.
7. Devore, J. L. *Probability and Statistics for Engineering and the Sciences* (Cengage learning, 2015).
8. Hogg, R. V., McKean, J. & Craig, A. T. Introduction to Mathematical Statistics, 7 ed (2012).
9. Pawitan, Y. *In all likelihood: statistical modelling and inference using likelihood* (Oxford University Press, 2001).
10. Lehmann, E. L. *Elements of large-sample theory* (Springer Science & Business Media, 1999).
11. Hoff, P. D. *A first course in Bayesian statistical methods* (Springer Science & Business Media, 2009).
12. Kay, S. M. Fundamentals of statistical signal processing, volume I: estimation theory (1993).
13. Ly, A., Marsman, M., Verhagen, J., Grasman, R. P. & Wagenmakers, E.-J. A tutorial on Fisher information. *Journal of Mathematical Psychology* **80,** 40–55 (2017).
14. Wilcox, R. R. *Fundamentals of modern statistical methods: Substantially improving power and accuracy* (Springer Science & Business Media, 2010).
15. Forbes, C., Evans, M., Hastings, N. & Peacock, B. *Statistical distributions* (John Wiley & Sons, 2011).
16. Krishnamoorthy, K. *Handbook of statistical distributions with applications* (CRC Press, 2016).