ISyE 3770: Introduction to Probability

In this section, we introduce some basic notions of probability. We will first introduce fundamental definitions, focusing on their motivations. These are followed by some concrete examples and exercises.

1 Random Experiments

Definition 1.1 (Random Experiment)**.** A random experiment is an experiment that can result in different outcomes, even though it is repeated in the same manner every time.

Example 1.1.

- Flipping a coin, rolling a dice.
- Lottery ticket winning number.
- Temperature tomorrow at noon.
- Drive time (minutes) from home to Georgia Tech at 8 am tomorrow.

To mathematically describe and then study a random experiment. There are several objects we next introduce. The first object we introduce is its sample space.

1.1 Sample Space of an Random Experiment

Definition 1.2 (Sample Space Ω)**.**

The set of all possible outcomes of a random experiment is called the sample space of the random experiment. We typically denote the sample space as Ω.

We use $\omega \in \Omega$ to say that ω belongs to the sample space Ω. That is, ω is a possible outcome of the random experiment.

[Graph illustration in Venn diagram]

- **Discrete sample space**: consists of a finite or countable infinite set of outcomes.
- **Continuous sample space**: contains an interval (either finite or infinite) of real numbers.

Example 1.2. Discrete sample space:

- Flipping a coin:

$$\text{A possible outcome } (\omega) : \text{Head}, \ \Omega = \{\text{Head, Tail}\}$$

- Rolling a dice:

$$\text{A possible outcome } (\omega) : 1, \ \Omega = \{1, 2, 3, 4, 5, 6\}.$$

Continuous sample space:

- Temperature tomorrow at noon:

$$\text{A possible outcome } (\omega) : 40, \ \Omega = \mathbb{R}.$$

- Drive time (minutes) from home to Georgia Tech at 8 am tomorrow:

$$\text{A possible outcome } (\omega) : 20, \ \Omega = \{x : x \geq 0\}.$$

2 Event of an Random Experiment

Having defined the sample space, we then define the event of a random experiments. Note that we can usually describe an event by just words. More formally, they are defined as follows.

Definition 2.1 (Event)**.** An event of a random experiment is a **subset** of the sample space (Ω). That is, any $E \subseteq \Omega$ is an event.

Remark 2.1. Note that an event $E \subseteq \Omega$ can contain multiple possible outcomes, or a single outcome. See examples below.

[Graph illustration in Venn diagram]

Example 2.1.

- Flipping a coin and get a tail:

$$E = \{\text{Tail}\}.$$

- Rolling a dice and either 1 or 2 faces up:

$$E = \{1, 2\}.$$

- Temperature tomorrow at noon ranges from 40 to 45 degrees:

$$E = [40, 45].$$

- Drive time (minutes) from home to Georgia Tech at 8 am tomorrow ranges from 5 to 15 minutes:

$$E = [5, 15].$$

3 Set (event) Operation

Since events are subsets of the sample space, we now discuss basic operations over events.

Definition 3.1 (Set Operations).

- **Union** of two events E_1 and E_2 is the event that consists of all outcomes that are contained in either of the two events.

$$E_1 \cup E_2 = \{\omega : \omega \in E_1 \text{ or } \omega \in E_2\}.$$

  [Graph illustration in Venn diagram]

- **Intersection** of two events is the event that consists of all outcomes that are contained in both of the two events.

$$E_1 \cap E_2 = \{\omega : \omega \in E_1 \text{ and } \omega \in E_2\}.$$

  [Graph illustration in Venn diagram]

- **Complement** of an event E in a sample space Ω is the set of outcomes in the sample space that are not in the event.

$$E^{\complement} = \{\omega : \omega \in \Omega \text{ but } \omega \notin E\}.$$

  [Graph illustration in Venn diagram]

Exercise 3.1. Suppose $\Omega = \{1, 2, 3, 4, 5\}$, $A = \{1, 2\}$, $B = \{1, 2, 3\}$. Then:

- $A \cup B$:
- $A \cap B$:
- $A^{\complement}$:

Solution:

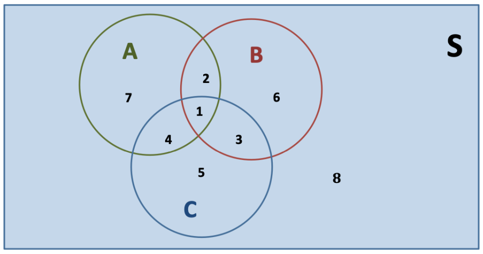
■

Definition 3.2 (Mutually Exclusive)**.** An empty set $\emptyset$ is a set that contains no element. Then E_1 and E_2 is mutually exclusive if

$$E_1 \cap E_2 = \emptyset.$$

[Graph illustration in Venn diagram]

Exercise 3.2. Consider the following Venn diagram. Here the number indicates the label of corresponding area in the Venn Diagram.



Questions:

- $C^{\complement}$:
- $B \cap C$:
- $A \cup C$:
- $B^{\complement} \cap A$:
- $A \cap B \cap C$:

- $(A \cup B) \cap C^{\complement}$:

Solution:

$\blacksquare$

4 Probability of an Random Experiment

As humans, it seems that we are intrinsically used to certain notions of frequency, based on our life experience. For instance, people living in Seattle might find it is more likely to rain in winter months of the year compared to summer. And if you roll a dice repeated for many times, the fraction of times you get 6 as the head will be approximately 1/6.

Now **probability**, as a mathematical tool, is invented in a way that **is consistent with our daily-life intuition**, and moreover, is **amenable to mathematical manipulation** so that we can use it to compute the likelihood of more complicated events. For instance,

What would be the likelihood of rolling a dice 100 times and observe that less than 30 times we obtain a number that is less than 4?

Of course, you can actually roll a dice that many times and count how many times that above even occurs. But that would be way too costly. This is exactly the reason we study probability more formally, so that we can analytically compute the likelihood of interest, without rolling a dice at all.

Definition 4.1 (Probability of a Random Experiment / Axiom of Probability)**.**

Probability is a function that assigns a number to any event $E \subset \Omega$ of a random experiment, denoted by $\mathbb{P}(E)$, which satisfies the following property:

1. $\mathbb{P}(\Omega) = 1$.
2. $0 \leq \mathbb{P}(E) \leq 1$, for any $E \subseteq \Omega$.
3. For two events E_1, E_2 with $E_1 \cap E_2 = \emptyset$, then

$$\mathbb{P}(E_1 \cup E_2) = \mathbb{P}(E_1) + \mathbb{P}(E_2).$$

Remark 4.1. It should be noted that probability is typically assigned or inferenced by human, as a reflection of our belief how the random experiment behaves.

For instance, when rolling a coin, we typically assign a probability of 1/2 for getting the head, as we *believe* the coin is fair.

Remark 4.2. Although events happen in a random fashion, the probability of these events happening is completely certain!

Exercise 4.1. Can an event E with zero probability (i.e., $\mathbb{P}(E) = 0$) happen? If so, can you give an example?

Solution:

■

Let us consider the following simple example of probability.

Definition 4.2 (Equally Likely Outcome)**.** Suppose Ω is discrete and contains N possible outcomes. That is, $\Omega = \{\omega_1, \dots, \omega_N\}$. Equally likely outcomes refers to a choice of probability $\mathbb{P}$ over Ω such that

$$\mathbb{P}(\{\omega_i\}) = \frac{1}{N},\ \forall \omega_i \in \Omega.$$

That is, every outcome has the same possibility.

Equally likely outcome as a choice of probability is very useful, with applications ranging from the simplest example to more complex ones.

Example 4.1. Flipping a coin:

$$\Omega = \{\text{Head, Tail}\}\,,\ \mathbb{P}(\{\text{Head}\}) = \mathbb{P}(\{\text{Tail}\}) = \frac{1}{2}.$$

Exercise 4.2. Suppose Ω is discrete and contains N possible outcomes, and the probability $\mathbb{P}$ produces equally likely outcomes. Let E be an event with k possible outcome. What is the probability of event E?

$$\mathbb{P}(E) = \frac{k}{N}.$$

Proof. This is intuitively very simple. But let us derive it from the definition (axiom) of probability and the definition of equally likely outcome.

□

Exercise 4.3. Suppose you roll a fair dice twice.

What is the probability that sum of the two trials equals to 2?

Solution:

■

Remark 4.3. As you can already see in the above example. Calculating probability of an even reduces to count the number of possible outcomes in the event, and the sample space, respectively. We will later discuss how to do counting efficiently.

4.1 Implication / Intuition of Axioms of Probability

We now discuss the implications of axioms of probability. Despite its simplicity, we will see that they are very intuitive. By using these simple rules, we can already calculate the probability of complex events.

Proposition 4.1. The axioms of probability imply that

1. $\mathbb{P}(\emptyset) = 0$ and $\mathbb{P}(E^{\complement}) = 1 - \mathbb{P}(E)$, for any $E \subseteq \Omega$.
2. If E_1 is contained in E_2, that is $E_1 \subseteq E_2$, then $\mathbb{P}(E_1) \leq \mathbb{P}(E_2)$.

Proof.

□

Example 4.2. Note that the second property above is very intuitive. For instance,

$$E_1 = \{\text{Tomorrow's temperature higher than 40 degrees}\},$$
$$E_2 = \{\text{Tomorrow's temperature higher than 50 degrees}\}.$$

Then as $E_2 \subseteq E_2$, we have $\mathbb{P}(E_1) \geq \mathbb{P}(E_2)$. This makes sense as whenever temperature is above 50 degrees, it also must exceed 40 degrees.

We now use axioms of probability to derive probabilities of more complex events.

Proposition 4.2. For any two events A and B, the probability of their union $A \cup B$ is given by:

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

If events A and B are mutually exclusive, $\mathbb{P}(A \cap B) = \mathbb{P}(\emptyset) = 0$. Then,

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

For the case of three events,

$$\mathbb{P}(A \cup B \cup C) = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) - \mathbb{P}(A \cap B) - \mathbb{P}(A \cap C) - \mathbb{P}(B \cap C) + \mathbb{P}(A \cap B \cap C).$$

If a collection of events $\{E_1, \ldots, E_N\}$ are pairwise mutually exclusive, i.e.

$$E_i \cap E_j = \emptyset, \forall i, j \in \{1, ..., N\}, i \neq j$$

then

$$\mathbb{P}(E_1 \cup E_2 \cup ... \cup E_N) = \sum_{i=1}^{N} \mathbb{P}(E_i).$$

Proof. Prove by Venn diagram. □

Exercise 4.4. Suppose $\Omega = A \cup B$, and $\mathbb{P}(A) = 0.8$, $\mathbb{P}(B) = 0.5$, then

$$\mathbb{P}(A \cap B) =$$

Solution:

■

Exercise 4.5. Suppose $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$, is A and B mutually exclusive?

Solution:

■

5 Counting Techniques and Probability

Definition 5.1 (Cardinality of Set)**.** The cardinality of a set E, denoted by $|E|$, is the number of items inside E.

What are counting techniques? Counts of the numbers of outcomes in the sample space (Ω) or event (E). That is, to calculate $|\Omega|$ and $|E|$.

When are they useful? They are particularly useful for calculating probability of event, in the discrete sample space (e.g., dice rolling). Recall Exercise 4.2, suppose Ω is discrete and contains N possible outcomes, and the probability $\mathbb{P}$ produces equally likely outcomes. Let E be an event with k possible outcomes, then the probability of event E is

$$\mathbb{P}(E) = \frac{k}{N} = \frac{|E|}{|\Omega|}.$$

Thus, we only need to compute $|E|$ and $|\Omega|$ in order to get the probability $\mathbb{P}(E)$.

A basic principle for counting techniques.

Proposition 5.1. If an operation can be described as a sequence of k steps and there are n_i ways of completing step i (for $i = 1, \ldots k$). Then the total number of ways of completing the operation is

$$n_1 \times n_2 \times \cdots \times n_k.$$

Below are several applications of Proposition 5.1.

Definition 5.2. A permutation of a set Ω is a unique ordered sequence of items inside Ω.

Example 5.1. Consider $\Omega = \{a, b, c\}$. Then all possible permutations are

$$(a, b, c), (a, c, b), (b, a, c), (b, c, a), (c, a, b), (c, b, a).$$

The total number of permutations are thus 6.

Proposition 5.2 (Counting Permutations of Distinct Items)**.** For a set Ω of n different items. The total number of possible permutations is

$$n! := n \times (n-1) \times \cdots \times 1.$$

By convention, we set $0! = 1$.

Proof. Can you construct a permutation from an n-step operation?

□

Exercise 5.1. There are 50 students in ISYE 3770 lecture room, with exactly 50 seats. Now suppose student coming in to the lecture room in order, and the ones coming earlier get to pick the seats freely whenever the seat is not taken.

What is the total number of possible seat arrangements for the classroom after everyone arrives?

Solution:

■

A slight generalization of the above is the selection while considering order.

Proposition 5.3 (Selection with Order)**.** Suppose Ω contains n distinct items. You select, at each time, an item from the remaining part of Ω. Suppose you select for a total of k times. The total number of possible combinations of the collected items is

$$P_k^n = n \times (n-1) \times \cdots \times (n-k+1) = \frac{n!}{(n-k)!}.$$

Proof. Consider a k-step operation.

□

Exercise 5.2. There are 50 students in ISYE 3770 lecture room. *Unfortunately, there are only 40 seats in total.* Now suppose student coming in to the lecture room in order, and the ones coming earlier get to pick the seats freely whenever the seat is not taken.

What is the total number of seat arrangements for the classroom when everyone arrives?

What is the probability that one particular student, Caleb, gets a seat?

Solution:

■

Exercise 5.3. Suppose every year is 365 days. There are 50 students in ISYE 3770. What is the probability that at least two students share the same birthday?

Solution:

■

Now suppose the set Ω has n items, *but some of them are the same.* Does it affect your answer on the total number of permutations?

Example 5.2. Consider $\Omega = \{a, b, b\}$, then all possible permutations are

$$(a, b, b), (b, a, b), (b, b, a).$$

The total number of permutations is 3.

More generally, we have the following.

Proposition 5.4 (Counting Permutations of When Some Items are the Same)**.** For a set Ω of n items. Suppose among the n items, there is

- n_1 Type-1 items that are identical,
- ...,
- n_k Type-k items that are identical.

The total number of possible permutations is

$$\frac{n!}{n_1! \times \cdots \times n_k!}.$$

Proof.

□

Exercise 5.4. Suppose there are 50 ISYE 3770 students coming from 10 different major, each major has exactly 5 students. There are 50 seats in total.

What is the total number of configurations when we look at the arrangement of seats in terms of students' majors?

Solution:

■

Proposition 5.5 (Combinations / Selection without Order / Selection without Replacement)**.** Suppose Ω contains n distinct items. You select, at each time, an item from the remaining part of Ω. Suppose you select for a total of k times. The total number of possible combinations, *without considering the order*, of the collected items is

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

Proof. Note the only difference between selection with and without order is whether we care about the order the selected k items. Since each possible outcome of selection without order corresponds to $k!$ possible outcome of selection with order, we have the total possible outcomes of selection with order given by

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

□

Exercise 5.5. Suppose there is 3-digit passcode (x, y, z). We know each of them is an integer, ranging from 1 to 9. We also know sum of them is 9.

What is the total number of possible passcodes?

Solution:

■

6 Conditional Probability

Recall that probability reflects how we believe each possible outcome takes place for a random experiment. For example, when tossing a coin, we believe

$$\mathbb{P}(\text{Getting a head}) = 1/2,$$

whenever we think the coin is fair.

Now, if we have some observations on this random experiment. In this case, it is quite natural to update our beliefs. That is, the observations we have on past trials of the random experiment give us extra information, and accordingly, we want to update the probability, based on these observations.

Example 6.1. Suppose you have a coin. You do not know if it is fair. But you know that for this coin, either it is fair (so that probability of a head is 0.5), or the probability of getting a head is 0.9.

Initially, it is natural to believe that both scenarios are equally likely, as you don't have any additional knowledge.

Suppose you have tossed the coin 1000 times, you observe there are 850 times of getting ahead.

Now, what would be your belief on whether the coin is fair?

Example 6.2. Consider a more trivial example. There are 3 balls in a bag, one green, two blue. Suppose you randomly pick a ball inside the bag (folding your eyes), and you observe that your pick is green.

Suppose you do not put the ball back. What is the probability that your next pick would be a blue ball? What about the probability of picking a green ball?

Conditional probability is a tool to help us quantify the above intuition.

Definition 6.1. Given an event A with $\mathbb{P}(A) > 0$, the conditional probability of an event B given event A, denoted as $\mathbb{P}(B|A)$, is defined as

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

Remark 6.1. It is very natural to see that in general,

$$\mathbb{P}(B|A) \neq \mathbb{P}(B).$$

To see this, consider tossing a fair coin,

$$B = \{\text{Getting the head}\}, \; A = \{\text{Getting the tail}\},$$

then we have

$$\mathbb{P}(B|A) = \frac{0}{1/2} = 0 \neq \mathbb{P}(B).$$

Remark 6.2. Definition 6.1 can be better understood as the following way. Instead of viewing the following

$$\mathbb{P}(B) \overset{\text{event A happens}}{\longrightarrow} \mathbb{P}(B|A),$$

which updates the probability of a single event B. We update the entire probability (belief):

$$\mathbb{P}(\cdot) \overset{\text{event A happens}}{\longrightarrow} \mathbb{P}(\cdot|A).$$

Proposition 6.1. Suppose the probability $\mathbb{P}$ is selected to satisfy equally likely outcome, then for any event A, we have

$$\mathbb{P}(A) = \frac{|\mathcal{A}|}{|\Omega|}.$$

Given this, we immediately see that for equally likely outcome,

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

Proof.

□

Example 6.3. Suppose we have a bag containing 5 items: 2 blue cube, 1 red cube, and 2 red balls. Now suppose we randomly sample from the bag, and observe that the shape is cube.

What is the probability the the sampled item is blue?

Solution:

■

Example 6.4. Suppose we have a bag, containing 2 blue balls, and 3 green balls. Consider the following scenarios.

(A). We sample from the bag 2 times. Each time, we select a ball randomly, observe the color, and put it back.

(B). We sample from the bag 2 times. Each time, we select a ball randomly, observe the color, and do not put it back.

Consider the event:

$$E = \{\text{getting 2 blue balls consecutively}\}.$$

Can you solve this using the notion of conditional probability instead of using the counting techniques?

Solution:

■

6.1 Intersection of Events

Sometimes, conditional probability is useful for calculating the probability of intersection of events.

Proposition 6.2. For any two events A and B, with $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$, then

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

Proof. By the definition of conditional probability. □

Example 6.5. Consider a truck traveling trying to from A to B, with a distance of 100 miles. For each mile, there is a probability of 0.1 that the truck broke down, provided that the truck hasn't broken down yet.

What is the probability that the truck finishes the trip?

Solution:

■

6.2 Total Probability Rules

Suppose an event A is pretty complex, and there is no simple way to directly evaluate its probability.

Proposition 6.3. For any event A, B, with $\mathbb{P}(A) > 0$, we have

$$\mathbb{P}(B) = \mathbb{P}(A) \cdot \mathbb{P}(B|A) + \mathbb{P}(A^{\mathsf{C}}) \cdot \mathbb{P}(B|A^{\mathsf{C}}).$$

Proof.

□

When would Proposition 6.3 be useful?

1. Event B is complex.
2. Event A and $A^{\complement}$ are simple.
3. Conditional probability of B given A and $A^{\complement}$ is easy.

Example 6.6. A manufacturing system depends on a component module M. Now if M works, then the system operates successfully with probability 0.95. If M does not work, then system operates successfully with probability 0.5.

Suppose M works with probability 0.8.

What is the probability that the system operates successfully?

Solution:

■

Example 6.7. Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, ”Do you want to pick door No. 2?” Is it to your advantage to switch your choice?

Solution: Consider two cases: if your initial selection is correct, and the initial selection is incorrect. What's the winning probability of each strategy under these two scenarios?

■

7 Independence

Recall that in general, the probability of event B conditioned on event A will not equal to the probability of event B,

$$\mathbb{P}(B|A) \neq \mathbb{P}(B), \text{ in general.}$$

If the equality indeed holds, this means that having the additional information on the event A happening does not affect our belief on how likely even B happens. This is formally called independence.

Definition 7.1. Event B and event A are independent if the following holds:

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

We have the following properties from the definition of conditional independence.

Proposition 7.1. If event A and event B are independent. Then

1. If $\mathbb{P}(A) > 0$, then $\mathbb{P}(B|A) = \mathbb{P}(B)$.
2. If $\mathbb{P}(B) > 0$, then $\mathbb{P}(A|B) = \mathbb{P}(A)$.

Proof.

□

Exercise 7.1. Questions:

1. If event A and event B are independent, are they mutually exclusive?

Solution:

■

Example 7.1. Consider you have a bag of 5 balls, 2 blue, 3 green. Consider the following scenarios:

1. You randomly select a ball, then put it back. Then randomly select the second ball.
2. You randomly select a ball, do not put it back. Then randomly select another ball from the remaining balls.

Consider

$$E_1 = \{\text{getting blue ball in the first pick}\},$$
$$E_2 = \{\text{getting blue ball in the second pick}\}$$

The question is:
(1) In Scenario 1, are E_1 and E_2 independent?
(2) How about for Scenario 2?

Solution:

■

Example 7.2. Suppose there are 2 paths from A to B. Each path has a probability of 0.2, that the road will be blocked by traffic jam. The traffic jams occur independently for these two roads.

What is the probability that one can travel from A to B without traffic jam?

Solution:

■

Example 7.3. Continuing from Example 7.2. Suppose there are 3 paths from B to C. Each path has a probability of 0.2 where traffic jam occurs. Again, traffic jams occur independently for these 5 roads.

There is no direct path from $A \to C$. The only way to travel from $A \to C$ is through $A \to B \to C$.

What is the probability that one can travel from A to C without traffic jam?

Solution:

■

Definition 7.2. Events $E_1, \ldots, E_n$ are independent if the following property holds: for any $1 \leq k \leq n$, and any $i_1, \ldots, i_k \in \{1, \ldots, n\}$, we have

$$\mathbb{P}(E_{i_1} \cap \cdots E_{i_k}) = \mathbb{P}(E_{i_1}) \cdot \cdots \cdot \mathbb{P}(E_{i_k}).$$

8 Bayes Theorem

Motivation: Bayes theorem is particularly helpful to update our prior belief once we obtain some new information.

Consider the example introduced at the beginning of Section 6.

Example 8.1. Suppose you have a coin. You do not know if it is fair. But you know that for this coin, only three cases can happen

1. fair: so that probability of a head is 0.5
2. slightly unfair: the probability of getting a head is 0.6
3. very unfair: the probability of getting a head is 0.9

Initially, you have not tossed the coin. It is natural to believe that both scenarios are equally likely to happen. That is

$$\mathbb{P}(\text{coin is fair}) = \mathbb{P}(\text{coin is slight unfair}) = \mathbb{P}(\text{coin is very unfair}) = 1/3.$$

Suppose you have tossed the coin 1000 times, and you observe there are 850 times of getting ahead. That is, you have observed that

$$B = \{850 \text{ heads out of 1000 independent trails}\} \text{ happens.}$$

To update our belief on the fairness of the coin, we are essentially asking

$$\mathbb{P}(\text{coin is fair} \mid B) \Rightarrow \text{ How would you evaluate this?}$$

Proposition 8.1 (Bayes Theorem)**.** Let $E_1, \dots, E_k$ be mutually exclusive, and

$$E_1 \cup E_2 \cup \cdots \cup E_k = \Omega,$$

and

$$\mathbb{P}(E_i) > 0, \text{ for any } 1 \leq i \leq k.$$

Then for any event B with $\mathbb{P}(B) > 0$, we have

$$\mathbb{P}(E_1|B) = \frac{\mathbb{P}(E_1)\mathbb{P}(B|E_1)}{\mathbb{P}(E_1) \cdot \mathbb{P}(B|E_1) + \cdots + \mathbb{P}(E_k) \cdot \mathbb{P}(B|E_k)}$$

Proof.

□

Exercise 8.1. Can we solve for the probability in Example 8.1 with Bayes theorem?

ISyE 3770: Discrete Random Variables

So far we have discussed the following key items for a random experiment:

- Sample space Ω: the set of all possible outcome
- Event $E \subset \Omega$: a subset of the sample space
- Probability $\mathbb{P}$: a function that assigns a number in $[0, 1]$ to every event

In addition, up to this point, when describing an outcome or an event, we are describing it by words. For example,

$$\text{E} = \text{flpping a coin 100 times and get 50 heads.}$$

In more complex setting, using words to describe an event/outcome becomes too complicated. A random variable is an alternative way to describe an event/outcome in such scenarios.

Definition 0.1. A random variable is a function that assigns a real number to each outcome in the sample space of a random experiment.

Example 0.1. Flip a coin 100 times. Let $\boldsymbol{X}$ = Total number of heads in this 100 trials. Then $\boldsymbol{X}$ is a random variable.

In addition, let $x = 50$, we can then see that

$$\{\boldsymbol{X} = x\} = \underbrace{\{\text{Flip a coin 100 times and observe 50 heads}\}}_{\text{Note that this is an event}}.$$

In general, for a random variable $\boldsymbol{X}$ and a possible value x, $\{\boldsymbol{X} = x\}$ typically denotes an event.

Again, we want to emphasize here that $\boldsymbol{X}$ itself is a random quantity. We only know that it takes certain value x with certain probability.

Definition 0.2. Depending on the precision of the random variable. We have

- Discrete random variable: there are only finitely many, or countably many possible values. For example, the $\boldsymbol{X}$ = number of heads when tossing a coin 100 times.
- Continuous random variable: has an interval (either finite or infinite) of real numbers as its range.
  For example, $\boldsymbol{X}$ = the temperature tomorrow at 12:00 pm.

In this chapter, we are solely focusing on discrete random variables.

1 Probability Distribution

First, it should be clear that for different value of x, the event $\{\boldsymbol{X} = x\}$ takes different probabilities.

The probability distribution of a random variable $\boldsymbol{X}$ is a description of the probabilities associated with the possible values of X. For a discrete random variable, the distribution is often specified by just a list of the possible values along with the probability of each.

Example 1.1. There is a chance that a bit transmitted through a digital transmission channel is received in error. Let X equal the number of bits in error in the next four bits transmitted. The possible values for X are 0, 1, 2, 3, 4. Suppose that the probabilities are

$$\begin{aligned}
&\mathbb{P}(\boldsymbol{X} = 0) = 0.6,\ \mathbb{P}(\boldsymbol{X} = 1) = 0.2 \\
&\mathbb{P}(\boldsymbol{X} = 2) = 0.1,\ \mathbb{P}(\boldsymbol{X} = 3) = 0.08 \\
&\mathbb{P}(\boldsymbol{X} = 4) = 0.02
\end{aligned}$$

The above specifies the probability distribution for random variable $\boldsymbol{X}$.

Example 1.2. Toss a coin three times. Let $\boldsymbol{X}$ denote the number of heads. What would be the probability distribution of $\boldsymbol{X}$?

Solution: Consider listing out all the possible outcome, and calculate the value of $\boldsymbol{X}$ for every outcome.

■

For discrete random variables, the probability distribution is also called probability mass function.

Definition 1.1 (Probability Mass Function)**.** For a discrete random variable $\boldsymbol{X}$ with possible value $x_1, \ldots, x_n$. Its probability mass function is a function $f(\cdot)$ such that

$$f(x_i) = \mathbb{P}(\boldsymbol{X} = x_i).$$

Of course, the probability mass function must satisfy

1. $f(x_i) \geq 0$, for any $i = 1, \ldots, n$.
2. $\sum_{i=1}^{n} f(x_i) = 1$.

Both of this follows directly from the axioms of probability we have discussed in Chapter 2.

Example 1.3. What is the probability mass function $f(\cdot)$ in the bits error example? (Example 1.1).

Solution:

■

Example 1.4. Consider a factory constantly producing bottled water. For each bottle, there is a 0.01 probability that the bottle contains a large particle of contamination. Suppose each bottle is independent. Once a contaminated water is produced, we can detect it immediately by some device.

Let $\boldsymbol{X}$ be the total number of bottled produced before we detect the first contaminated bottle. Question: What is the probability mass function for $\boldsymbol{X}$?

Solution:

■

Below is an example that might be helpful to adjust a common misconception when we first encounter probability mass function.

Example 1.5. Suppose $\boldsymbol{X}$ is a random variable, taking values in $\{1, 2, 3, ...\}$. Let $f(\cdot)$ denote its probability mass function. Can the following happen?

$$f(i) > 0, \text{ for all } i = 1, 2, 3, ...$$

Solution: Consider the following function

$$f(i) = \frac{1}{i \cdot (i+1)}.$$

Is it a valid probability mass function?

■

Definition 1.2 (Cumulative Distribution)**.** The cumulative distribution of a random variable $\boldsymbol{X}$, denoted as $F(\cdot)$, is a function given by

$$F(x) = \mathbb{P}(\boldsymbol{X} \leq x) = \sum_{x_i \leq x} f(x_i).$$

Example 1.6. What would be the cumulative distribution of the random variable $\boldsymbol{X}$ in Example 1.1?

Solution: Write out the answer.

Draw the graph for $F(\cdot)$. Observe that there are jumps in the graph.

■

Example 1.7. Why is cumulative distribution useful?

Solution:

1. It can provide direct answer to what is the probability of $\{\boldsymbol{X} \leq x\}$.
2. It can help determine the probability mass function.

Why does Claim 2 hold?

In short, this is because

$$f(x_i) = F(x_i) - \lim_{x \uparrow x_i} F(x).$$

■

Exercise 1.1. Here is some basic properties of a cumulative distribution.

1. $0 \leq F(x) \leq 1$.
2. For any $x \leq y$, we have $F(x) \leq F(y)$.

Can we show this?

Solution:

■

For discrete random variables, probability mass function is typically more useful for many applications.

2 Expectation and Variance of Discrete Random Variables

The expectation of $\boldsymbol{X}$ quantifies, on average, what would be the value of $\boldsymbol{X}$ (note that the $\boldsymbol{X}$ itself is random). On the other hand, the variance of $\boldsymbol{X}$ quantifies, how large does $\boldsymbol{X}$ deviates from its expectation (average).

Example 2.1. Consider the temperature in your fridge, versus the temperature of your balcony.

The fridge is constantly, consistently much colder. While the balcony can be cold, hot, breezy, warm, etc. On average, balcony temperature is higher, and with much larger variation (uncertainty).

Definition 2.1 (Mean and Variance)**.** The mean, or the expected value of the discrete random variable $\boldsymbol{X}$, denoted as μ or $\mathbb{E}(\boldsymbol{X})$, is

$$\mu = \mathbb{E}[\boldsymbol{X}] = \sum_x x f(x).$$

Here $f(\cdot)$ denotes the probability mass function of $\boldsymbol{X}$. In addition, the variance of $\boldsymbol{X}$, denoted as σ^2 or $V(\boldsymbol{X})$, is

$$\sigma^2 = V(\boldsymbol{X}) = \mathbb{E}(\boldsymbol{X} - \mu)^2 = \sum_x (x - \mu)^2 f(x) = \sum_x x^2 f(x) - \mu^2.$$

The standard deviation of $\boldsymbol{X}$ is defined as $\sigma = \sqrt{\sigma^2}$.

In other words, mean can be viewed as the weighted sum of all possible values, where the weight for a given value x is the probability of $\boldsymbol{X}$ taking value x (i.e., $f(x)$).

Variance can be viewed as the weighted sum of the squared distance from each value x_i to the mean μ, where the weight for each x is again $f(x)$.

Exercise 2.1. Let $f(\cdot)$ be the probability mass function of random variable $\boldsymbol{X}$. Let μ denote the mean of $\boldsymbol{X}$. Why does the following holds?

$$\sum_x (x-\mu)^2 f(x) = \sum_x x^2 f(x) - \mu^2.$$

Solution:

■

Example 2.2. We know that σ^2 roughly quantifies how much $\boldsymbol{X}$ deviates from its mean. See the following graph illustration of probability mass function where the expectation is fixed as 0 but we can increase the variance.

Now what does it mean if $\sigma^2 = 0$?

Solution:

■

Let us see for a concrete example where we can use mean and variance to compare different random variables.

Example 2.3. Two investment products, both with the same cost to purchase. Let $\boldsymbol{X}$ denote the return of the first product. $\boldsymbol{Y}$ denote the return of the second product.

Consider the following two cases:

1. The probability mass function of $\boldsymbol{X}$ is

$$f_{\boldsymbol{X}}(1) = 0.5, f_{\boldsymbol{X}}(0) = 0.5.$$

Meanwhile, the probability mass function of $\boldsymbol{X}$ is

$$f_{\boldsymbol{Y}}(10) = 0.05, f_{\boldsymbol{X}}(0) = 0.95.$$

2. The probability mass function of $\boldsymbol{X}$ is

$$f_{\boldsymbol{X}}(-1) = 0.5, f_{\boldsymbol{X}}(1) = 0.5.$$

Meanwhile, the probability mass function of $\boldsymbol{X}$ is

$$f_{\boldsymbol{Y}}(2) = 0.5, f_{\boldsymbol{X}}(0) = 0.5.$$

What would be your choice of product to invest in each of the cases?

Solution: Let us compute the mean and the variance for these two cases.

■

Example 2.4. Suppose $\mathbb{E}(\boldsymbol{X}) = \mathbb{E}(\boldsymbol{Y})$, and $V(\boldsymbol{X}) = V(\boldsymbol{Y})$, does it mean that $\boldsymbol{X} = \boldsymbol{Y}$?

Solution: Consider $\boldsymbol{X}$ with the following probability mass function

$$f(-1) = 0.5, \ f(10) = 0.05.$$

Can you construct another random variable $\boldsymbol{Y}$ with the same expectation and variance?

■

The previous example show that the mean and the variance does not uniquely determine a random variable. Two different variable can share the same mean and variance. In some sense, mean and variance only roughly describe a random variable, but the information summarized by these two quantities are not rich enough.

In this cases, comparing two different variables (for instance, return of investment) becomes tricky. We need to design some other metric that can capture additional information.

In many cases, this reduce to compute the expectation of *some function of the random variable.*

Proposition 2.1. Let $\boldsymbol{X}$ be a random variable with probability mass function $f(\cdot)$. Then for any function $h(\cdot)$, we have

$$\mathbb{E}[h(\boldsymbol{X})] = \sum_x h(x)f(x).$$

Example 2.5. Consider the investment example (Example 2.3). What is $\mathbb{E}(\boldsymbol{X}^3)$ in the first case?

Solution:

■

Below, we discuss a property of expectation and variance that will be extremely useful in daily applications, even after the semester ends.

Proposition 2.2 (Mean and Variance of Linear Function of Random Variable)**.** For any random variable $\boldsymbol{X}$, any $a, b \in \mathbb{R}$, we have

$$\begin{aligned}\mathbb{E}(a\boldsymbol{X} + b) &= a\mathbb{E}(\boldsymbol{X}) + b,\\ V(a\boldsymbol{X} + b) &= a^2V(\boldsymbol{X}).\end{aligned}$$

Proof. Can we just show it by the definition of expectation and variance?

Can we use this property to immediately apply to some previous example?

For instance, what is $V(-\boldsymbol{X})$? (see Example 2.4).

□

3 Discrete Uniform Distribution

This is perhaps the simplest discrete random variable.

Definition 3.1. A random variable $\boldsymbol{X}$ has a discrete uniform distribution if each of the n values in its range, $x_1, \ldots, x_n$, has equal probability. That is

$$f(x_i) = \frac{1}{n}, \text{ for all } i = 1, \ldots, n.$$

Below let us determine the mean and variance of discrete uniform random variable when its range is $\{a, a+1, \ldots, b\}$, where $a \leq b$ and both are integers.

Proposition 3.1. Let $\boldsymbol{X}$ has the discrete uniform distribution, with range $\{a, a+1, \ldots, b\}$, where $a \leq b$ and both are integers. Then we have

$$\mu = \mathbb{E}(\boldsymbol{X}) = \frac{a+b}{2}, \sigma^2 = V(\boldsymbol{X}) = \frac{(b-a+1)^2 - 1}{12}.$$

Proof.

□

Exercise 3.1. Now suppose for some integer $n \geq 1$. $\boldsymbol{Y}$ is the discrete random variable with uniform distribution, with range from $\{n \cdot a, n \cdot (a+1), \ldots, n \cdot b\}$, where $a \leq b$ and both are integers. Questions:

$$\mathbb{E}(\boldsymbol{Y}) = ?, \ V(\boldsymbol{Y}) = ?.$$

Solution: Hint: What can we say about the relation between $\boldsymbol{X}$ and $\boldsymbol{Y}$?

■

4 Binomial Distribution

Let us first introduce the concept of Bernoulli trial.

Definition 4.1. A random experiment is called a Bernoulli trial if it only has two possible outcomes. We typically label one of these outcomes as "success" and the other one as "failure". Of course, whichever outcome you would like to call a "success" is totally up to you. The key parameter of a Bernoulli trial is the probability of "success", denoted as p, that is:

$$p = \mathbb{P}(\text{success}),\ 1 - p = \mathbb{P}(\text{failure}).$$

Some examples are in order.

Example 4.1. The following are Bernoulli trials:

1. Tossing a coin: "success" means getting a head, "failure" means getting the tail.

   What is p here?

2. Transmitting a code where error happens with probability 0.01: "success" means no transmission error, "failure" means transmission with an error.

   What is p here?

3. Randomly guesses an answer for a 2-choice question: "success" means correct guess, "failure" means incorrect guess.

   What is p here?

Definition 4.2. We can now define the so-called Bernoulli random variable. For a Bernoulli trial with "success" probability $0 < p < 1$, $\boldsymbol{X}$ is called a Bernoulli random variable with parameter p if

$$\boldsymbol{X} = \begin{cases} 1, & \text{if trial is "success"}, \\ 0, & \text{if trial is "failure"}. \end{cases}$$

The immediate question on Bernoulli random variables.

Exercise 4.1. What is the expectation and the variance of a Bernoulli random variable with parameter p?

Solution:

■

With the Bernoulli trial in place, we can now discuss Binomial distribution.

Definition 4.3. Consider a random experiemnt consisting of n Bernoulli trials, and

1. The trials are independent.
2. Each trial results in only two possible outcomes, labeled as "success" and "failure."
3. The probability of a success in each trial, denoted as p, remains constant.

Let $\boldsymbol{X}$ = the total number of "success" trials. We then call $\boldsymbol{X}$ a Binomial random variable with parameters $0 < p < 1$ and n.

The probability mass function of $\boldsymbol{X}$ is given by

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \ x = 0, 1, \ldots, n. \tag{4.1}$$

Example 4.2. Going back to our 4-digit code transmission problem (Example 1.1). Suppose each digit is transmitted with an error probability of 0.01. Let $\boldsymbol{X}$ = total number of digits transmitted with error. Could you give the probability mass function of $\boldsymbol{X}$?

Solution: Hint: What is the probability for a possible outcome that contains exactly x-digits error? And what is the total number of outcomes that contains exactly x-digits error?

■

Going one-step further.

Exercise 4.2. Why does the probability mass function is given (4.1)?

Solution: Hint: What is the probability for a possible outcome that contains exactly x "success" trials? And what is the total number of outcomes that contains exactly x "success" trials?

■

Exercise 4.3. Can you draw and visualize the probability mass function of Binomial random variable with parameters p and n? What is the most salient trait for this graph?

Solution: Hint: where do we reach the maximum approximately? Is it always symmetric?

■

Proposition 4.1. The expectation and the variance of Binomial random variable $\boldsymbol{X}$ with parameters p and n is given by

$$\mathbb{E}(\boldsymbol{X}) = np,\ V(\boldsymbol{X}) = np(1-p).$$

Proof. Chapter 5 will show us how to compute the variance easily.

Here let us focus on showing $\mathbb{E}(\boldsymbol{X})$, how would you approach it?

Hint: Consider $\boldsymbol{X} = \boldsymbol{X}_1 + \boldsymbol{X}_2 + \cdots + \boldsymbol{X}_n$, how would you describe $\boldsymbol{X}_1$, $\boldsymbol{X}_2$ ... $\boldsymbol{X}_n$? Can it be some Bernoulli random variables?

□

Exercise 4.4. What would be the mean and the variance for $\boldsymbol{X}$ in the 4-digt code transimission problem (Example 4.2)?

Solution:

■

Exercise 4.5. Now define $\boldsymbol{Y} = \boldsymbol{X}/n$. Then $\boldsymbol{Y} =$ The frequency of "successful" trials.

What would you say about the mean and the variance of $\boldsymbol{Y}$? Moreover, what happens if n goes to infinity?

Solution: Hint: Can you see how variance behaves if n goes to infinity? Can you link this behavior to Example 2.2?

■

5 Geometric Distribution

A practical example to introduce the geometric distribution.

Example 5.1. Suppose we flip a coin repeatedly. Let X be the total number of flips until we observe the head for the first time. What is the probability mass function of $\boldsymbol{X}$?

Solution: Hint: what is the probability that $\boldsymbol{X} = 1, 2, \ldots k, \ldots$?

■

Since flipping a coin is an example of a Bernoulli trial. The geometric distribution is formally described as follows.

Definition 5.1. Consider repeatedly conducting a series of independent Bernoulli trial with parameter $0 < p < 1$. Let $\boldsymbol{X}$ be the number of trials until the first success. Then $\boldsymbol{X}$ is called a geometric random variable with parameter p, with its probability mass function given by

$$f(x) = (1-p)^{x-1} \cdot p, \;\; x = 1, 2, \dots$$

Example 5.2. Consider the water contamination example (Example 1.4), where each produced bottle has a 0.01 probability of being contaminated. What is the probability that exactly 120 bottles need to be examined until we detect the first contaminated bottle?

Solution:

■

Proposition 5.1. The mean and the variance of the geometric random variable $\boldsymbol{X}$ with parameter p is given by

$$\mathbb{E}\left[\boldsymbol{X}\right] = \frac{1}{p}, \;\; V(\boldsymbol{X}) = \frac{1-p}{p^2}.$$

Proof. For simplicity, let us just focus on deriving $\mathbb{E}\left[\boldsymbol{X}\right]$.

□

Example 5.3. What is the mean and the variance of the total number of bottles we have to exam to detect the first contaminated bottle? (Example 1.4)

How large is the standard deviation σ compared to the mean μ? What can you conclude?

Why do we want to compare standard deviation σ with μ? Why not comparing the variance σ^2 with mean μ?

Solution:

■

5.1 Lack-of-Memory Property of Geometric Distribution

This is perhaps the most important property of the geometric distribution.

Example 5.4. Again, consider the water contamination example. Suppose for the first 100 bottles, we have detected no contaminated water. What is the probability that we detect the first contaminated bottle at the 110-th bottle?

Solution: Hint: can we desribe the answer as the conditional probability? Can you derive the answer by following the definition of conditional probability?

■

Example 5.5. Let us define $\boldsymbol{X}$ as the total number of bottled water examined to until we detect the first contaminatd bottle? Question: Can you describe the conditional probability above using just notation of $\boldsymbol{X}$?

Solution:

■

As you can see from the above example, conditioned on the event that $\{\boldsymbol{X} \geq n+1\}$, the probability of $\boldsymbol{X} = n+k$, is given exactly by $\mathbb{P}(\boldsymbol{X} = k)$. That is, if the "success" (detecting contaminated water) still hasn't happened at the n-th bottle (i.e., $\boldsymbol{X} \geq n+1$), then the probability of success happening at the $(n+k)$-th bottle is the same as if: we restart the examination process and detect the first contaminated bottle at the k-th bottle.

Definition 5.2. Suppose a random variable $\boldsymbol{X}$ take values in $1, 2, \ldots$. The $\boldsymbol{X}$ is said to satisfies lack-of-memory (or memoryless) property if

$$\mathbb{P}(\boldsymbol{X} = n+k | \boldsymbol{X} > n) = \mathbb{P}(\boldsymbol{X} = k), \ \forall n = 1, 2, \ldots, \ \forall k = 1, 2, \ldots$$

Example 5.6. Can you verify that the geometric distribution with parameter p satisfies the memoryless property stated in Definition 5.2?

Solution:

■

Example 5.7. Do you think geometric variable is the only type of discrete random variables that satisfy lack-of-memory property?

Solution: Hint: the answer at first appears to be fairly surprising. This is also I believe why geometric distribution receives so much attention among all the discrete random variables.

■

6 Poisson Distribution

Poisson distribution is perhaps one of the most widely used distributions to describe events that randomly occur over an interval of time or space.

Consider the following example.

I would say that the application of Poisson random variables, identifying in which case you can use them, is perhaps best illustrated in the following example where we introduce the concept of the Poisson process.

Example 6.1. Let T denote the length of a wire in millimeters. Suppose λ is the expected (averaged) number of flaws per millimeter. Let $\boldsymbol{X}$ denote the total number of flaws on this wire.

We can model the distribution of $\boldsymbol{X}$ in the following way.

1. Divide the wire into n segments.

2. We assume that on each segment, we either have zero, or one flaw. (note that here we are doing an approximation argument, this is reasonable as n becomes very large). That is, the number of errors on the k-th segment, can be modeled into a Bernoulli random variable with parameter p.

3. What is p here? Note that
$$\mathbb{E}\left[\boldsymbol{X}\right] = \lambda \cdot T = n \cdot p,$$
we have $p = \lambda \cdot T/n$.

4. Hence $\boldsymbol{X}$, as the total number of flaws on the n segments, can be modeled into a Binomial distribution, with parameter (n, p).

5. We then know that the distribution of $\boldsymbol{X}$ has an approximate probability mass function as
$$\begin{aligned} f(x) &= \binom{n}{x} p^x (1-p)^{n-x} \\ &= \binom{n}{x} \left(\frac{\lambda T}{n}\right)^x \left(1 - \frac{\lambda T}{n}\right)^{n-x}, \ x = 0, 1, \ldots, n. \end{aligned}$$

6. **The last step, what would be the distribution of $\boldsymbol{X}$ look like if $n \to \infty$?** That is, what if we take the segment to be finer and finer?

Solution:

■

Generalizing the above process in the previous example, let us define the Poisson process.

Definition 6.1. A random experiment is called a Poisson process if the following holds.

1. There is an interval (either in time or space) of T
2. The interval can be divided into sub-intervals, each of length Δt, which goes to 0
3. The probability of one "success" on the sub-interval tends $\lambda \cdot \Delta t$, and the probability of two or more "success" on the sub-interval tends to 0
4. Event on the sub-intervals is independent of each other

Definition 6.2. The random variable $\boldsymbol{X}$ that denotes the number of "successes" is called a Poisson random variable with parameter (λ, T), with its probability mass function given by

$$f(x) = \frac{(\lambda T)^x}{x!} \exp^{-\lambda T}, \ x = 0, 1, 2, \ldots .$$

Let us see now for a concrete application of Poisson distribution.

Example 6.2. Suppose that the number of flaws follows a Poisson distribution with a mean of 2.3 flaws per millimeter.

1. What is the probability of 10 flaws in 5 millimeters of wire?
2. What is the probability of at least one flaw in 2 millimeters of wire?

Solution: Hint: can 10 flaws occur on 5 millimeters of wire?

What does the mean of 2.3 flaws per millimeter decide about the Poisson random variable here?

■

Now let us determine the mean and the variance of a Poisson random variable.

Proposition 6.1. The mean and the variance of a Poisson random variable with parameters (λ, T) is given by

$$\mathbb{E}[\boldsymbol{X}] = \lambda T, \ V(\boldsymbol{X}) = \lambda T.$$

Proof. We can use the following useful trick:

$$\sum_{x=1}^{\infty} \frac{(\lambda T)^{x-1}}{(x-1)!} = \partial_y \left(\sum_{x=1}^{\infty} \frac{y^x}{x!} \right) \Bigg|_{y=\lambda T} = \partial_y (\exp(y) - 1) \Bigg|_{y=\lambda T} = \exp(\lambda T).$$

□

Example 6.3. If a random variable $\boldsymbol{X}$ has a significantly different mean and variance. Would it be a good choice to model it as a Poisson random variable?

Solution: Hint: how do the mean and variance of Poisson distribution compare to each other?

■

ISyE 3770: Continuous Random Variables

As we have discussed in the previous chapter. The layout of this chapter will be similar. First, we are going to discuss probability density, cumulative distribution function, the expectation, and the variance for general continuous random variables. Then we are going to talk about some concrete and useful examples of continuous random variables. These includes the uniform distribution, the normal (Gaussian) distribution, and the exponential distribution.

Definition 0.1. A continuous random variable $\boldsymbol{X}$ is a random variable with an interval (either finite or infinite) of real numbers for its range.

1 Probability Density Function

For convenience, we can always assume that the range of $\boldsymbol{X}$ is $\mathbb{R} := (-\infty, \infty)$.

Definition 1.1. For a continuous random variable $\boldsymbol{X}$, its probability density function is a function f such that

1. $f(x) \geq 0$, for any $x \in \mathbb{R}$.
2. $\int_{-\infty}^{\infty} f(x)\ dx = 1$.
3. The probability of $\boldsymbol{X}$ is fully specified in the following sense. For any $A \subset \mathbb{R}$, we have

$$\mathbb{P}(\boldsymbol{X} \in A) = \int_A f(x)\ dx.$$

   In particular, we have for any $a \leq b$,

$$\mathbb{P}(a \leq \boldsymbol{X} \leq b) = \int_a^b f(x)\ dx.$$

The following exercise should help us understand the difference between the probability mass function of a discrete random variable and the probability density function of a continuous random variable.

Exercise 1.1. Consider the following questions:

1. Can the probability mass function $f(x)$ takes value that is strictly greater than 1? That is, can $f(x) = 2$ happen? What if f is a probability density function?

2. If $\boldsymbol{X}$ is a continuous random variable, what about

$$\mathbb{P}(\boldsymbol{X} = x) = ?$$

Can we also state the following?

$$\mathbb{P}(a < \boldsymbol{X} \leq b) = \mathbb{P}(a \leq \boldsymbol{X} < b) = \mathbb{P}(a < \boldsymbol{X} < b) = \mathbb{P}(a \leq \boldsymbol{X} \leq b)?$$

Solution:

■

Given the answer to the second question above, why is the probability mass function useful at all? It it only gives you a trivial answer to $\mathbb{P}(\boldsymbol{X} = x)$?

Consider the following example.

Example 1.1. Consider $\boldsymbol{X}$ and two scenarios.

1. The probability density function is given by

$$f(x) = \begin{cases} 0.5, & 0 \leq x \leq 2, \\ 0, & \text{otherwise}. \end{cases}$$

2. The probability density function is given by

$$f(x) = \begin{cases} 0.8, & 0 \leq x \leq 1, \\ 0.2, & 1 < x \leq 2, \\ 0, & \text{otherwise}. \end{cases}$$

Now what is the following probability under these two scenarios?

$$\mathbb{P}(1 - \epsilon \leq \boldsymbol{X} \leq 1) = ?, \text{ where } 0 < \epsilon < 1.$$

Can you see how the value of f close to 1 changes the answer to the above probability?

Solution:

■

In summary, the area the probability density function under the area $[x-\epsilon, x+\epsilon]$ decides the probability of $\{x-\epsilon \leq \boldsymbol{X} \leq x+\epsilon\}$.

It would be best to understand the value of $f(x)$ as the likelihood of $\{\boldsymbol{X} \approx x\}$, instead of the probability of $\{\boldsymbol{X} = x\}$.

More complicated example to illustrate the full power of a probability density function in determining the probability of complex events.

Example 1.2. Suppose the density function of a random variable $\boldsymbol{X}$ is given as follows:

$$f(x) = \begin{cases} e^{-x}, \ x \geq 0, \\ 0, \ \text{otherwise}. \end{cases}$$

First, verify that f is a valid probability density function.

Now, can you also determine the following probability?

$$\mathbb{P}(0 \leq \boldsymbol{X} \leq 0.5, \ \text{or } 1 \leq \boldsymbol{X} \leq 1.5, \ \text{or } \cdots, k \leq \boldsymbol{X} \leq k+0.5, \ \text{or } \cdots)$$

Solution:

■

2 Cumulative Distribution

Similar to the discrete case, the cumulative distribution of a continuous random variable $\boldsymbol{X}$ is given as follows.

Definition 2.1. The cumulative distribution of a continuous random variable $\boldsymbol{X}$ is

$$F(x) = \int_{-\infty}^{x} f(x)\ dx,$$

where f is the density function of $\boldsymbol{X}$. Note that by definition we have

$$F(x) = \mathbb{P}(\boldsymbol{X} \leq x).$$

From the definition, once given the probability density function, we can decide the cumulative distribution function.

An example illustrating how one can obtain the cumulative distribution given the density function.

Example 2.1. Suppose the density function of $\boldsymbol{X}$ is given by

$$f(x) = \begin{cases} 0.8, & 0 \leq x \leq 1, \\ 0.2, & 1 < x \leq 2, \\ 0, & \text{otherwise.} \end{cases}$$

What is the cumulative distribution F?

Solution:

■

Now what is we are given the cumulative distribution function $F(x)$, and tries to recover the density function $f(x)$?

Proposition 2.1. Let F denote the cumulative distribution function of continuous random variable $\boldsymbol{X}$. Then the density function is simply given by

$$f(x) = F'(x),\ \forall x \in \mathbb{R},$$

as long as the gradient $F'(x)$ exists for x.

Proof.

□

Example 2.2. Can you apply the above proposition to Example 2.1? What would you say when $x = 0$ or $x = 1$ or $x = 2$?

Solution:

■

Next, we show that the cumulative distribution function $F(x)$ is as useful as the density function $f(x)$, when trying to determine the probability of events involving random variable $\boldsymbol{X}$.

Proposition 2.2. Suppose continuous random variable $\boldsymbol{X}$ has a cumulative distribution function F. Let $a \leq b$, then

$$\mathbb{P}(a \leq \boldsymbol{X} \leq b) = F(b) - F(a).$$

Moreover, we also have

$$\mathbb{P}(\boldsymbol{X} \geq a) = 1 - F(a).$$

Proof.

□

The following example summarizes what we have discussed in this section.

Example 2.3. Suppose continuous random variable $\boldsymbol{X}$ has a cumulative distribution function F given by

$$F(x) = \begin{cases} 0, \ x \leq 0, \\ 1 - e^{-0.01x}, \ x > 0. \end{cases}$$

Can you determine the density function f? In addition, what is the following probability

$$\mathbb{P}(\boldsymbol{X} \leq 100) = ?, \ \mathbb{P}(\boldsymbol{X} > 200) = ?$$

Solution:

■

3 Mean and Variance of Continuous Random Variables

The mean and variance for continuous random variables are essentially, defined in exactly the same way as the discrete random variables.

Definition 3.1. The mean $\mathbb{E}[\boldsymbol{X}]$ is defined as

$$\mu = \mathbb{E}[\boldsymbol{X}] = \int_{-\infty}^{\infty} x f(x) \ dx.$$

The variance $V(\boldsymbol{X})$ is defined as

$$\sigma^2 = V(\boldsymbol{X}) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) \ dx = \int_{-\infty}^{\infty} x^2 f(x) - \mu^2.$$

Again, the standard deviation is defined as $\sigma = \sqrt{\sigma^2}$.

Exercise 3.1. Can we show the last equality of the definition of $V(\boldsymbol{X})$?

$$\int_{-\infty}^{\infty} (x - \mu)^2 f(x) \ dx = \int_{-\infty}^{\infty} x^2 f(x) - \mu^2?$$

Solution:

■

Example 3.1. Suppose the probability density function of $\boldsymbol{X}$ is given by

$$f(x) = \begin{cases} 0, \ x < 0, \\ \exp(-x), \ x \geq 0. \end{cases}$$

What are the expectation and the variance of $\boldsymbol{X}$?

Solution:

■

We can also define the expectation of complex function applied to $\boldsymbol{X}$.

Proposition 3.1. Let $\boldsymbol{X}$ have a probability density function f. Let h be a function mapping from $\mathbb{R}$ to $\mathbb{R}$, and define $\boldsymbol{Y} = h(\boldsymbol{X})$. Then

$$\mathbb{E}\left[\boldsymbol{Y}\right] = \mathbb{E}\left[h(\boldsymbol{X})\right] = \int_{-\infty}^{\infty} h(x) f(x) \ dx.$$

Example 3.2. Let the density function of $\boldsymbol{X}$ be given by

$$f(x) = \begin{cases} 0, \ x < 0, \\ \exp(-x), \ x \geq 0, \end{cases}$$

what is the expectation of $\boldsymbol{Y} = \exp\left(\frac{1}{2}\boldsymbol{X}\right)$?

Solution:

■

The following is a very interesting result, at least when I first saw it, that can sometimes help directly compute the expectation from the **cumulative** distribution function of $\boldsymbol{X}$.

Proposition 3.2. Let $\boldsymbol{X}$ be a continuous random variable with $\boldsymbol{X} \geq 0$ (that is, $f(x) = 0$ for all $x < 0$) Then the expectation of $\boldsymbol{X}$ is also given by

$$\mathbb{E}\left[\boldsymbol{X}\right] = \int_0^\infty (1 - F(x))\, dx.$$

Let us first see an immediate application of Proposition 3.2.

Exercise 3.2. Suppose continuous random variable $\boldsymbol{X}$ has a cumulative distribution function F given by

$$F(x) = \begin{cases} 0, \; x \leq 0, \\ 1 - e^{-0.01x}, \; x > 0. \end{cases}$$

What is the expectation $\mathbb{E}\left[\boldsymbol{X}\right]$? Can you use two approaches? One from the definition, and the other one from Proposition 3.2?

Solution: Which approach do you think is simpler?

■

Exercise 3.3. Can you prove Proposition 3.2? In fact, can you show that more general claim?

$$\mathbb{E}\left[\boldsymbol{X}^p\right] = \int_0^\infty px^{p-1}(1 - F(x))\ dx.$$

Solution:

■

4 Continuous Uniform Distribution

Definition 4.1. A continuous random variable $\boldsymbol{X}$ with a probability density function

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b; \\ 0, & \text{otherwise} \end{cases}$$

is called a continuous uniform random variable over $[a, b]$.

It is relatively straightforward to compute the expectation and the variance of uniform distribution.

Proposition 4.1. If $\boldsymbol{X}$ is a continuous uniform random variable over $[a, b]$. Then

$$\mathbb{E}\left[\boldsymbol{X}\right] = \frac{a+b}{2};\ V(\boldsymbol{X}) = \frac{(b-a)^2}{12}.$$

Proof.

□

It is also quite natural to obtain the cumulative distribution function of uniform distribution. Consider the following exercise.

Exercise 4.1. If $\boldsymbol{X}$ is a continuous uniform random variable over $[a, b]$. Then what is the cumulative distribution function of $\boldsymbol{X}$?

Solution:

■

5 Normal/Gaussian Distribution

Before introducing the concrete definition, it is worth highlighting that the Gaussian distribution is perhaps the most widely used, and hence the most important distribution in probability and statistics.

To see how Gaussian distributions naturally arise. Consider the following example.

Example 5.1. Consider the following two scenarios.

1. Suppose $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ are independent discrete Bernoulli random variable, with equal mean μ and equal variance σ^2
2. Suppose $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ are independent continuous Uniform random variables over $[0, 1]$, with equal mean μ and variance σ^2

Let us consider

$$\boldsymbol{Y} := \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i - \mu$$

What do we know from our previous chapter?

We know that $\boldsymbol{Y}$ converge to the constant 0 as n goes to infinity.

Question: what do you think would be

$$\sqrt{n} \cdot \boldsymbol{Y}/\sigma, \text{ when } n \text{ goes to infinity?}$$

Solution: This is known as the Central Limit Theorem, which tells us in both scenarios, $\sqrt{n}\boldsymbol{Y}$ will have a Gaussian distribution $N(0, 1)$.

Why do I say it is surprising?

$\boldsymbol{X}_i$'s are completely different random variables in Scenarios 1 and 2, but their "scaled average" has the same distribution!

■

The above example shows that the "scaled average" of a sequence of independent random variables typically has a Gaussian distribution – this is also the reason why Gaussian distribution arises so frequently in natural systems. In many scenarios the random variable of interest can be viewed as the aggregation of many small random effects – and hence can be viewed as the "scaled average" of such effects.

Definition 5.1. A random variable $\boldsymbol{X}$ with probability density function

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

is called a normal/Gaussian random variable with parameters $-\infty < \mu < \infty$ and $\sigma^2 > 0$.

Moreover, we have

$$\mathbb{E}\left[\boldsymbol{X}\right] = \mu,\ V(\boldsymbol{X}) = \sigma^2.$$

We typically use $N(\mu, \sigma^2)$ to denote the normal distribution.

A unique feature of the normal distribution is that it is very light tailed.

Example 5.2. Can you draw the shape of the density function of $N(\mu, \sigma^2)$? What can you say about the tail of the density function?

Solution:

Indeed, we have the following

$$\begin{aligned}\mathbb{P}(\mu - \sigma \leq \boldsymbol{X} \leq \mu + \sigma) &= 0.68\\ \mathbb{P}(\mu - 2\sigma \leq \boldsymbol{X} \leq \mu + 2\sigma) &= 0.95\\ \mathbb{P}(\mu - 3\sigma \leq \boldsymbol{X} \leq \mu + 3\sigma) &= 0.99.\end{aligned}$$

As you can see, the normal distribution basically assigns zero probability for any values outside the interval $[\mu - 3\sigma, \mu + 3\sigma]$. ■

How do we get the above observation on the tail of $N(\mu, \sigma^2)$, do we have to integration for every possible (μ, σ)?

In fact, we can first just focus on the standard normal distribution.

Definition 5.2. A normal random variable with

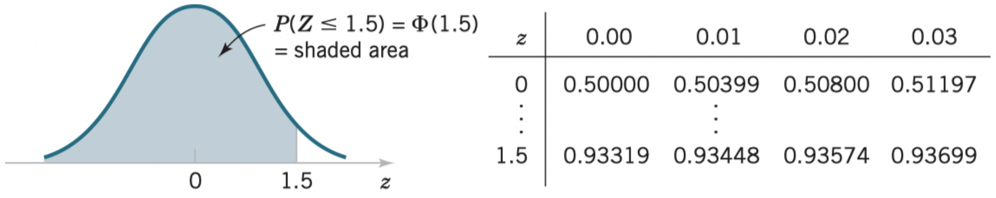
$$\mu = 0, \sigma = 1,$$

is called a standard normal distribution, typically denoted as $\boldsymbol{Z}$. We will also use a special notation for its cumulative distribution function

$$\Phi(z) = \mathbb{P}(\boldsymbol{Z} \leq z).$$

It should be noted that $\Phi(z)$ is not easy to be calculated. For this purpose, we typically precompute $\Phi(z)$ for some values of z and summarize them in a table. This table is typically called a Φ-table, which is very useful, and we illustrate its usage as follows.

Example 5.3. Below is a part of the Φ-table. The table provides $\Phi(z) = \mathbb{P}(\boldsymbol{Z} \leq z)$ as follows.



z	0.00	0.01	0.02	0.03
0	0.50000	0.50399	0.50800	0.51197
⋮		⋮		
1.5	0.93319	0.93448	0.93574	0.93699

Solution:

■

Of course, once you have $\Phi(z)$, you can compute a lot of different things

Exercise 5.1. Using the table above, can you compute:

$$\mathbb{P}(0 \leq \boldsymbol{Z} \leq 1.5) = ?; \ \mathbb{P}(\boldsymbol{Z} \geq 0.01) = ?$$

Solution:

■

We will see that for any probability questions on arbitrary $N(\mu, \sigma^2)$, we can answer it by using knowledge on $N(0, 1)$.

Proposition 5.1. If $\boldsymbol{X}$ is a normal random variable with mean μ and variance σ^2. Then

$$\frac{\boldsymbol{X} - \mu}{\sigma}$$

is a standard normal random variable, with mean 0 and variance 1.

Consequently, for any $\boldsymbol{X}$ with a distribution $N(\mu, \sigma^2)$, we typicall call the transformation

$$\boldsymbol{Z} = \frac{\boldsymbol{X} - \mu}{\sigma}$$

as standardizing the random variable $\boldsymbol{X}$.

Now why is Proposition 5.1 useful in any sense? Consider the following example.

Example 5.4. Suppose that the current measurements in a strip of wire, denoted as $\boldsymbol{X}$ are assumed to follow a normal distribution with a mean of 10 milliamperes and a variance of 4 (milliamperes) 2 . What is the probability that a measurement exceeds 13 milliamperes?

Solution:

■

The above procedure can be summarized as in the following proposition.

Proposition 5.2. Consider any normal random variable $\boldsymbol{X}$ with distribution $N(\mu, \sigma^2)$. Define

$$\boldsymbol{Z} = \frac{\boldsymbol{X} - \mu}{\sigma}.$$

For any $-\infty < x < \infty$, let us also define

$$z = \frac{x - \mu}{\sigma}.$$

Then we have

$$\mathbb{P}(\boldsymbol{X} \leq x) = \mathbb{P}(\frac{\boldsymbol{X} - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}) = \mathbb{P}(\boldsymbol{Z} \leq z) = \Phi(z),$$

and the last quantity can be easily found with a Φ-table.

Example 5.5. Under the same setup as in Example 5.4, what is

$$\mathbb{P}(9 \leq \boldsymbol{X} \leq 10) \text{ expressed in terms of } \Phi(z)?$$

Solution:

■

6 Normal Appproximation to Binomial and Poisson Distributions

6.1 Normal Approximation to Binomial Distribution

Example 6.1. I claim that in the Scenario 1 Example 5.1, we effectively show that if $\boldsymbol{X}$ follows a Binomial distribution with parameters (n, p), then we have

$$\frac{\boldsymbol{X} - np}{\sqrt{np(1-p)}}$$

follows approximately the standard normal distribution $N(0, 1)$.

Question: why is the above statement true?

Solution: Hint: can you decompose $\boldsymbol{X}$ into a summation of independent Bernoulli random variables?

■

Now why do we care about using the normal approximation?

Exercise 6.1. Suppose we are transmitting 16 million digits, and each digit got received with an error independently, with an error 10^{-5}.

What is the probability that there are 150 or fewer errors?

Solution: Hint: what is the distribution of $\boldsymbol{X}$ – the total number of errors? Can you derive the exact formula for the probability of interest? Can you actually compute it exactly?

If exact computation is not possible, is there any reasonable approximation scheme?

■

The following proposition provides a general procedure to approximate the probability of Binomial distribution with parameter (n, p) with a normal distribution, provided the n is large enough.

Proposition 6.1. If $\boldsymbol{X}$ is a Binomial random variables with parameters (n, p), then

$$\boldsymbol{Z} = \frac{\boldsymbol{X} - np}{\sqrt{np(1-p}}$$

is an approximately standard normal random variable. That is, we have the following approximation:

$$\mathbb{P}(\boldsymbol{X} \leq x) = \mathbb{P}(\boldsymbol{X} \leq x + 0.5) = \mathbb{P}(\boldsymbol{Z} \leq \frac{x + 0.5 - np}{\sqrt{np(1-p}}) \approx \Phi(\frac{x + 0.5 - np}{\sqrt{np(1-p}}).$$

Moreover,

$$\mathbb{P}(\boldsymbol{X} \geq x) = \mathbb{P}(\boldsymbol{X} \geq x - 0.5) = \mathbb{P}(\boldsymbol{Z} \geq \frac{x - 0.5 - np}{\sqrt{np(1-p}}) \approx 1 - \Phi(\frac{x - 0.5 - np}{\sqrt{np(1-p}}).$$

In general, teh approximation $\approx$ above will be accurate for $np > 5$ and $n(1-p) > 5$.

Example 6.2. Consider Example 6.1, can you provide a solution using the normal approximation procedure?

Solution:

■

To see how accurate the approximation is, let us consider the following example.

Example 6.3. Consider transmitting 50 bits, each with a probability of error 0.1. What is the probability that 2 or fewer errors occur, considering using both the exact solution and the normal approximation?

In addition, what is the probability that exactly 5 errors occur, considering using both the exact solution and the normal approximation?

Solution:

■

6.2 Normal Approximation to Poisson Distribution

Let us start directly from the proposition.

Proposition 6.2. If $\boldsymbol{X}$ is a Poisson random variable with expectation λ and variance also λ. Then

$$\boldsymbol{Z} := \frac{\boldsymbol{X} - \lambda}{\sqrt{\lambda}},$$

is an approximate standard normal random variable.

In general, the approximation is accurate if $\lambda > 5$.

Proof. Here can we recall how do we motivate the definition of a Poisson random variable? (Consider $T = 1$).

Reminder: consider a random variable $\boldsymbol{X}$ that has a Binomial distribution with parameters (n, p), if $n \to \infty$, and $np \to \lambda$, then $\boldsymbol{X}$ will converge to a Poisson random variable with mean λ as n goes to infinity.

□

Example 6.4. Assume that the number of asbestos particles in a squared meter of dust on a surface follows a Poisson distribution with a mean of 1000. If a squared meter of dust is analyzed, what is the probability that 950 or fewer particles are found?

Solution: Hint: Can you first derive the formula for the exact probability? Do you see the difficulty if you want to compute the exact number?

How would you approach it with an approximation argument?

■

7 Exponential Distribution

The exponential distribution, as a continuous distribution, is intimately connected with the discrete Poisson distribution! Consider the following example.

Example 7.1. Let the total number of flaws on the copper wire follow a Possion distribution with a rate of $\lambda > 0$ (that is, λ flaws per millimeter). Let $\boldsymbol{X}$ denote the length of the wire starting from one flaw until we get the next flaw. What would be the distribution of $\boldsymbol{X}$?

Solution: Hint: Suppose $\boldsymbol{X} \geq x$, then what does it mean?

Does this mean that for the x millimeters of wire, there is no flaw detected?

What should be the number of flaws for this x millimeter of wire in the first place?

■

Definition 7.1. Consider Poisson process that has a mean of $\lambda > 0$ events per unit interval. Let random variable $\boldsymbol{X}$ be the distance between successive events in such a process. Then $\boldsymbol{X}$ is called an exponential random variable with parameter $\lambda > 0$. Then density function is given by

$$f(x) = \begin{cases} \lambda \exp(-\lambda x), \ x \geq 0; \\ 0, \ \text{otherwise}. \end{cases}$$

Proof of PDF. See the above example for the proof of the probability density function. □

We now proceed to establish the mean and the variance of exponential distributions.

Proposition 7.1. Let $\boldsymbol{X}$ be an exponential random variable with parameter λ, then

$$\mathbb{E}[\boldsymbol{X}] = \frac{1}{\lambda}; \ V(\boldsymbol{X}) = \frac{1}{\lambda^2}.$$

Proof.

□

7.1 Lack of Memory Property of Exponential Random Variable

Consider the following example that illustrates the notion of “lack of memory” property.

Example 7.2. Let $\boldsymbol{X}$ denotes the time between customer arrivals and assume that $\boldsymbol{X}$ follows exponential distribution with $\mathbb{E}[\boldsymbol{X}] = 1.4$. Consider the following two scenarios:

1. Suppose we just opened the counter, what is the probability that a customer arrives in the next 30 seconds?
2. Suppose the counter has seem no arrival of customer for 3 minutes. What is the probability that a new customer arrives at the counter in the next 30 seconds?

It seems that in the second scenario, we have waited for a long time, so it seems more likely that we get a new customer with higher change. But is this intuition true?

Solution: Hint: can we formalize the second scenario into a conditional probability?

■

The above example is indeed a special case of the following general claim.

Proposition 7.2. Suppose $\boldsymbol{X}$ is an exponential random variable, then

$$\mathbb{P}(\boldsymbol{X} < t_1 + t_2 | \boldsymbol{X} > t_1) = \mathbb{P}(\boldsymbol{X} < t_2).$$

Proof.

□

Do you think the lack of memory property is surprising? If so, why is that? Below, let us consider a explanation of why this lack of memory property is not surprising.

The following example is also a good summary of the relation between Binomial distribution and the Poisson distribution.

Example 7.3. Let us recall the following fact, which occurs in the beginning of our discussion of Poisson process in the previous chapter.

Consider a Binomial distribution with parameters (n, p). Let $n \to \infty$, and $p \to 0$, while $np = \lambda$. Then the probability mass function of the Binomial distribution with parameters (n, p) converges to the probability mass function of the Poisson random variable with mean λ.

Graphically speaking, we are doing nothing but finer-and-finer division of a length-1 interval into n sub-intervals, and a "success event" occurring on each sub-interval with probability λ/n, and the "success event" on different sub-intervals are independent of each other.

Given the above construction of a Poisson distribution from a limit of Binomial distribution, we can immediately see that

At any sub-interval, the history on the previous sub-intervals (that is, whether "success" happens or not) has absolutely no effect on whether "success" happens at this interval.

This explains why the gap between two "successes" in the Poisson process has no memory. Consequently, the exponential random variable has no memory.

Now let us go one more step. Do you think there is any other continuous distribution that satisfies the lack-of-memory property?

Proposition 7.3. The exponential distribution is the only continuous distribution that satisfies the lack-of-memory property.

Proof.

□

ISyE 3770: Joint Probability Distributions

Before we start any technical discussions, I would like to highlight that this chapter is perhaps the most important chapter for the probability content in this course. Recall that in previous chapters we have discussed discrete and continuous random variables. Our attention was paid to a **single** random variable, studying its origin, its mass/density function, and its applications. Now this chapter help us understand and formalize how different random variables interact with each other.

Putting it in another way, previous chapters provide us raw materials for describing the **simple random events** in daily life. It is well likely that for realistic random events, we might not find a proper probability distribution that we have discussed and can describe this event.

Now in this chapter, we are going to discuss how we can put those raw materials together to actually build complicated random variables that are much more powerful, and help us model **rich/complex random events** that we are trying to model.

1 Joint Distribution for Two Random Variables

To motivate, let us consider the following example for a discrete random variable.

Example 1.1. Suppose $\boldsymbol{X}$ and $\boldsymbol{Y}$ are both Bernoulli random variables with parameter $p \in (0, 1)$. Then what is the probability that $\boldsymbol{X} = \boldsymbol{Y}$?

Solution: Hint: do you think this problem is well-defined or not?

■

Consider the next example for a continuous random variable.

Example 1.2. Suppose $\boldsymbol{X}$ is a standard normal random variable. Let $\boldsymbol{Y} = -\boldsymbol{X}$, what is the distribution of $\boldsymbol{Y}$? Let $\boldsymbol{Z} = \boldsymbol{X}$.

Now consider the following probability:

$$\mathbb{P}(\boldsymbol{X} \geq 0, \boldsymbol{Y} \geq 0) = ?; \ \mathbb{P}(\boldsymbol{X} \geq 0, \boldsymbol{Z} \geq 0) = ?$$

Solution:

■

The previous example shows that for two random variables or multiple random variables in general. It is very important to explicitly specify their joint behavior. Specifying their individual behavior is simply not enough to determine the probability of outcomes that involves both of them.

In general, if X and Y are two random variables, the probability distribution that defines their simultaneous behavior is called a joint probability distribution.

1.1 Joint Probability Mass Function

If both $\boldsymbol{X}$ and $\boldsymbol{Y}$ are discrete random variables, we can define the joint probability mass function that determines their joint behavior.

Definition 1.1 (Joint Probability Mass Function)**.** Suppose $\boldsymbol{X}$ take possible values in discrete set $\mathcal{X}$, $\boldsymbol{Y}$ takes possible values in discrete set $\mathcal{Y}$. The joint probability mass function of $(\boldsymbol{X}, \boldsymbol{Y})$, denoted by $f_{\boldsymbol{XY}}(x, y)$ is defined as

$$f_{\boldsymbol{XY}}(x, y) = \mathbb{P}(\boldsymbol{X} = x, \boldsymbol{Y} = y).$$

Consequently, this function satisfies

$$\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} f_{\boldsymbol{XY}}(x, y) = 1; \ f_{\boldsymbol{XY}}(x, y) \geq 0, \text{ for any } x \in \mathcal{X}, y \in \mathcal{Y}.$$

Typically, a joint probability mass function can be simply represented by a table. Here is a simple application.

Example 1.3. Consider the following table that determines the joint probability mass function of $(\boldsymbol{X}, \boldsymbol{Y})$, both are Binomial random variables with parameter $p = 1/2$.

Compute the probability that $\boldsymbol{X} = \boldsymbol{Y}$.

Solution:

$\blacksquare$

1.2 Joint Density Function

If both $\boldsymbol{X}$ and $\boldsymbol{Y}$ are continuous random variables, we can define the joint probability density function that determines their joint behavior.

Definition 1.2 (Joint Probability Density Function)**.** Suppose $\boldsymbol{X}, \boldsymbol{Y}$ are continuous random variables taking values in $\mathbb{R}$, The joint probability density function of $(\boldsymbol{X}, \boldsymbol{Y})$, denoted by $f_{\boldsymbol{XY}}(x, y)$ is a function satisfying

1. Non-negativity: $f_{\boldsymbol{XY}}(x, y) \geq 0$, for any $(x, y) \in \mathbb{R}^2$.
2. Unit integral:
$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{\boldsymbol{XY}}(x, y)\mathrm{d}x\mathrm{d}y = 1.$$
3. Consistent with probability:
$$\mathbb{P}((\boldsymbol{X}, \boldsymbol{Y}) \in \mathcal{R}) = \int\int_{(x,y)\in\mathcal{R}} f_{\boldsymbol{XY}}(x, y)\mathrm{d}x\mathrm{d}y.$$

Remark 1.1. It should be noted that in view of the Condition 3 of Definition 1.2, a jointy density function fully specifies the joint behavior of $(\boldsymbol{X}, \boldsymbol{Y})$. Moreover, the probability $\mathbb{P}((\boldsymbol{X}, \boldsymbol{Y}) \in \mathcal{R})$ can be understood as the "volume under the surface", where the "surface" is the value of function $f(x, y)$ over the region $\mathcal{R}$. See the figure below.

To showcase the power of joint density function, let us consider the following example.

Example 1.4. Let $(\boldsymbol{X}, \boldsymbol{Y})$ has a joint density function given by

$$f_{\boldsymbol{XY}}(x, y) = \begin{cases} 1, & 0 \le x \le 1,\ 0 \le y \le 1, \\ 0, & \text{otherwise.} \end{cases}$$

Now

- Verify that the above is a valid joint density function.
- Compute the following probability:

$$\begin{aligned} \mathbb{P}(0 \le \boldsymbol{X} \le \frac{1}{2}, 0 \le \boldsymbol{Y} \le \frac{1}{2}) &= ?; \\ \mathbb{P}(\boldsymbol{X} \ge \boldsymbol{Y}) &= ?; \\ \mathbb{P}(\boldsymbol{Y} \ge \boldsymbol{X}^2) &= ?. \end{aligned}$$

Solution:

■

1.3 Marginal Distribution

The previously discussed joint density function $f_{\boldsymbol{XY}}(x, y)$ specifies the joint behavior of random variables $(\boldsymbol{X}, \boldsymbol{Y})$. What we have not emphasized is that once we have specified the joint density function, we have indeed also determined the individual behavior of each random variable $\boldsymbol{X}$ and $\boldsymbol{Y}$.

We first introduce the notion of marginal probability mass function for discrete random variables.

Definition 1.3 (Marginal Probability Mass Function)**.** Let $(\boldsymbol{X}, \boldsymbol{Y})$ be discrete random variables with a joint mass function of $f_{\boldsymbol{XY}}(x, y)$, with $\boldsymbol{X}$ taking values in a discrete set $\mathcal{X}$, and $\boldsymbol{Y}$ taking values in a discrete set $\mathcal{Y}$. Then the marginal probability mass function of $\boldsymbol{X}$ and $\boldsymbol{Y}$ are defined, respectively, as

$$f_{\boldsymbol{X}}(x) = \sum_{y \in \mathcal{Y}} f_{\boldsymbol{XY}}(x, y);$$

$$f_{\boldsymbol{Y}}(y) = \sum_{x \in \mathcal{X}} f_{\boldsymbol{XY}}(x, y).$$

Moreover, the marginal density function $f_{\boldsymbol{X}}(x)$ governs the behavior of random variable $\boldsymbol{X}$, and marginal density function $f_{\boldsymbol{Y}}(y)$ governs the behavior of random variable $\boldsymbol{Y}$, in the sense that,

$$\mathbb{P}(\boldsymbol{X} = x) = f_{\boldsymbol{X}}(x); \tag{1.1}$$

$$\mathbb{P}(\boldsymbol{Y} = y) = f_{\boldsymbol{Y}}(y). \tag{1.2}$$

Exercise 1.1. Question: why does the marginal mass function satisfies properties (1.1) and (1.2)?

Solution: Hint: can you start from the definition of the joint probability mass function and the marginal mass function?

■

Definition 1.4 (Marginal Probability Density Function)**.** Let $(\boldsymbol{X}, \boldsymbol{Y})$ be continuous random variables with a joint density function of $f_{\boldsymbol{XY}}(x, y)$. Then the marginal probability density function of $\boldsymbol{X}$ and $\boldsymbol{Y}$ are defined, respectively, as

$$f_{\boldsymbol{X}}(x) = \int_{-\infty}^{\infty} f_{\boldsymbol{XY}}(x, y) \mathrm{d}y;$$

$$f_{\boldsymbol{Y}}(y) = \int_{-\infty}^{\infty} f_{\boldsymbol{XY}}(x, y) \mathrm{d}x.$$

Moreover, the marginal density function $f_{\boldsymbol{X}}(x)$ governs the behavior of random variable $\boldsymbol{X}$, and marginal density function $f_{\boldsymbol{Y}}(y)$ governs the behavior of random variable $\boldsymbol{Y}$, in the sense that,

$$\mathbb{P}(a \leq \boldsymbol{X} \leq b) = \int_a^b f_{\boldsymbol{X}}(x) \mathrm{d}x; \tag{1.3}$$

$$\mathbb{P}(a \leq \boldsymbol{Y} \leq b) = \int_a^b f_{\boldsymbol{Y}}(y) \mathrm{d}y. \tag{1.4}$$

Exercise 1.2. Question: why does the marginal density satisfies properties (1.3) and (1.4)?

Solution: Hint: can you start from the definition of the joint probability density function and the marginal density function?

■

Of course, given the joint mass/density function, one can use them to compute the expectation and the variance of each random variable.

Proposition 1.1. Let $(\boldsymbol{X}, \boldsymbol{Y})$ be discrete random variables with joint probability mass function as $f_{\boldsymbol{XY}}(x, y)$, then the expectation, the variance of $\boldsymbol{X}$ is given by

$$\mu_{\boldsymbol{X}} = \mathbb{E}\left[\boldsymbol{X}\right] = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} x f(x, y),$$
$$\sigma^2_{\boldsymbol{X}} = \mathbb{E}\left[(\boldsymbol{X} - \mu_{\boldsymbol{X}})^2\right] = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} (x - \mu_{\boldsymbol{X}})^2 f(x, y).$$

The expectation and the variance of $\boldsymbol{Y}$ can be computed similarly.

A similar proposition also holds for continuous random variables.

Proposition 1.2. Let $(\boldsymbol{X}, \boldsymbol{Y})$ be continuous random variables with joint probability density function as $f_{\boldsymbol{XY}}(x, y)$, then the expectation, the variance of $\boldsymbol{X}$ is given by

$$\mu_{\boldsymbol{X}} = \mathbb{E}\left[\boldsymbol{X}\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f(x, y)\mathrm{d}x\mathrm{d}y,$$
$$\sigma^2_{\boldsymbol{X}} = \mathbb{E}\left[(\boldsymbol{X} - \mu_{\boldsymbol{X}})^2\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x - \mu_{\boldsymbol{X}})^2 f(x, y)\mathrm{d}x\mathrm{d}y.$$

The expectation and the variance of $\boldsymbol{Y}$ can be computed similarly.

Proof. Hint: the proof follows directly from the expectation and the variance of $\boldsymbol{X}$ we have discussed in previous chapters, together with (1.1), (1.3). □

Example 1.5. Let the joint density function of $(\boldsymbol{X}, \boldsymbol{Y})$ be given as

$$f_{\boldsymbol{XY}}(x, y) = \begin{cases} 6 \times 10^{-6} \exp(-0.001x - 0.002y), \ 0 \le x < y; \\ 0, \ \text{otherwise}. \end{cases}$$

Compute the expectation of $\boldsymbol{X}$.

Solution:

■

2 Conditional Probability Distributions and Independence

When we were previously saying "the relation between random variables $\boldsymbol{Y}$ and $\boldsymbol{Y}$", one particular question on their relationship can be framed as follows

Suppose we know that $\boldsymbol{X} = x$, how should we accordingly change/update our belief on how the random variable $\boldsymbol{Y}$ behaves?

In particular, since we typically express our "belief" on the behavior of $\boldsymbol{Y}$ through its density/mass function, ths leads to the notion of "conditonal mass/density function".

Definition 2.1 (Conditional Mass Function)**.** Let $(\boldsymbol{X}, \boldsymbol{Y})$ be discrete random variables with a joint mass function of $f_{\boldsymbol{X}\boldsymbol{Y}}(x, y)$, with $\boldsymbol{X}$ taking values in a discrete set $\mathcal{X}$, and $\boldsymbol{Y}$ taking values in a discrete set $\mathcal{Y}$. Then the conditional mass function of $\boldsymbol{Y}$ conditioned on $\{\boldsymbol{X} = x\}$ is defined by

$$f_{\boldsymbol{Y}|x}(y) = \frac{f_{\boldsymbol{X}\boldsymbol{Y}}(x, y)}{f_{\boldsymbol{X}}(x)}, \text{ for any } x \text{ with } f_{\boldsymbol{X}}(x) > 0.$$

Importantly, the conditional mass function satisfies the following property:

$$f_{\boldsymbol{Y}|x}(y) = \mathbb{P}(\boldsymbol{Y} = y|\boldsymbol{X} = x). \tag{2.1}$$

Exercise 2.1. Why does the conditional mass function satisfies the property (2.1)?

Proof. Hint: proof by definition.

□

Now turning our attention to the continuous random variables, how do we define the notion of "conditional probability density function"?

Definition 2.2. Let $(\boldsymbol{X}, \boldsymbol{Y})$ be continuous random variables with joint probability density function as $f_{\boldsymbol{X}\boldsymbol{Y}}(x, y)$. Then the conditional density function of $\boldsymbol{Y}$ given $\{\boldsymbol{X} = x\}$ is defined as

$$f_{\boldsymbol{Y}|x}(y) = \frac{f_{\boldsymbol{X}\boldsymbol{Y}}(x, y)}{f_{\boldsymbol{X}}(x)}, \quad \text{for any } x \text{ such that } f_{\boldsymbol{X}}(x) > 0.$$

Remark 2.1. It is important to note that we can only condition on $\{\boldsymbol{X} = x\}$ if $f_{\boldsymbol{X}}(x) > 0$, otherwise the conditional density function $f_{\boldsymbol{Y}|x}$ is not well defined.

The previous definition only tells us the expression for the conditional density function, but what properties does it have? Are these properties intuitive?

Proposition 2.1. We have the following holds.

1. $f_{\boldsymbol{Y}|x}(\cdot)$ itself is a valid probability density function. (But it is different than $f_{\boldsymbol{Y}}(\cdot)$!) That is,

$$f_{\boldsymbol{Y}|x}(y) \geq 0,\ \forall y \in \mathbb{R};\ \int_{-\infty}^{\infty} f_{\boldsymbol{Y}|x}(y)\mathrm{d}y = 1.$$

2. It holds that

$$\mathbb{P}(a \leq \boldsymbol{Y} \leq b|\boldsymbol{X} = x) = \int_a^b f_{\boldsymbol{Y}|x}(y)\mathrm{d}y.$$

Importantly, we should note that given the above properties the conditional density function is a valid density function.

Now the first property should be understood in this way: whenver we have known $\boldsymbol{X} = x$, then the density function (our belief on $\boldsymbol{Y}$) shoud be updated from the original marginal density of $\boldsymbol{Y}$ – that is, $f_{\boldsymbol{Y}}(\cdot)$, to another density function, which is $f_{\boldsymbol{Y}|x}(\cdot)$.

On the other hand, the second property above itself is non-trivial. Indeed, let us ask ourself a related question. If $\boldsymbol{X}$ is a continuous random variable, the $\boldsymbol{X} = x$ happens with ZERO probability! How do we define the condition probability?

$$\mathbb{P}(a \leq \boldsymbol{Y} \leq b|\boldsymbol{X} = x)?$$

Is it obvious?

Below is a pretty intuitive definition, that is a result a taking the limit of a sequence of well-defined conditional probability.

Definition 2.3. For continuous random variables $\boldsymbol{X}$ and $\boldsymbol{Y}$, we define

$$\mathbb{P}(a \leq \boldsymbol{Y} \leq b|\boldsymbol{X} = x) = \lim_{\epsilon \to 0, \epsilon \neq 0} \mathbb{P}(a \leq \boldsymbol{Y} \leq b|x - \epsilon \leq \boldsymbol{X} \leq x + \epsilon).$$

Without going into measure-theory course, the following argument is the most typically used thought process for deriving (understanding) the conditional density function. This also shows that conditional density function is defined via the limitation of a certain operation.

Example 2.1. Consider $\epsilon > 0$, and a pair of continuous random variables $(\boldsymbol{X}, \boldsymbol{Y})$. Could you derive

$$\mathbb{P}(a \le \boldsymbol{Y} \le b \mid x \le \boldsymbol{X} \le x + \epsilon) = ?$$

Could you give an approximation of the solution above and relate it to the second property of Proposition 2.1?

Solution:

■

In daily life, if we are trying to use conditional density, we sometimes need to pay extra attention. Let us consider in the following example, a quite famous paradox.

Example 2.2 (Borel–Kolmogorov paradox)**.** We will parameterize the earth's surface: let (θ, λ) be longitude and latitude of a traveler, respectively (in radians), where $0 \le \theta \le 2\pi$ and $-\frac{\pi}{2} \le \lambda \le \frac{\pi}{2}$. Suppose the traveller's location is uniformly distributed on the earth surface. This can be done by specifying the joing distribution of (θ, λ) as

$$f(\theta, \lambda) = \begin{cases} \frac{1}{4\pi} \cos\lambda, & 0 \le \theta \le 2\pi, \quad -\frac{\pi}{2} \le \lambda \le \frac{\pi}{2}; \\ 0, & \text{otherwise.} \end{cases}$$

Now consider the following scenario:

1. Suppose we are given the information that our random explorer is on the equator. What is the conditional distribution of the longitude given that latitude is 0?

   Answer: uniform distribution.

   $$f_{\boldsymbol{\Theta}|\lambda}(\theta) = \frac{1}{2\pi}, \ 0 \le \theta \le 2\pi$$

   Makes sense right? The location is uniformly distributed

2. Suppose we are given the information that our random explorer is on the prime meridian (0 degrees longitude), which is 1/2 of a great circle (the equator is an entire great circle). What is the conditional distribution of the latitude given that longitude is 0?

   Answer: cosine distribution.

   $$f_{\boldsymbol{\Lambda}|\theta}(\lambda) = \frac{1}{2}\cos(\lambda), \ -\frac{\pi}{2} \le \lambda \le \frac{\pi}{2}$$

   Paradox: Why is not uniform? We essentially have the same circle.

Solution: Hint: both solutions are perfectly correct. We just have to interpret the conditional probabilities above following the procedure of Example 2.1.

■

Now let us turn our attention back to simpler examples.

Example 2.3. Let the joint density function of $(\boldsymbol{X}, \boldsymbol{Y})$ be given as

$$f_{\boldsymbol{XY}}(x, y) = \begin{cases} 6\exp(-x-2y), & 0 \leq x < y; \\ 0, & \text{otherwise.} \end{cases}$$

Compute the conditional density functions $f_{\boldsymbol{X}|y}(x)$ and $f_{\boldsymbol{Y}|x}(y)$.

Solution:

■

How do we understand the differences between marginal mass/density function $f_{\boldsymbol{X}}$ and conditional mass/density function $f_{\boldsymbol{X}|y}$?

Marginal mass/density function: Originally, without any additional information, we use the marginal mass/density function $f_{\boldsymbol{X}}$ to represent our **belief** on the behavior of random variable $\boldsymbol{X}$.

Conditional mass/density function: Now suppose we have obtained the **extra information** that $\{\boldsymbol{Y} = y\}$, then we use the marginal mass/density function $f_{\boldsymbol{X}|y}$ to represent our **refreshed belief** on the behavior of random variable $\boldsymbol{X}$.

In general, marginal and conditional density/mass functions are completely different!

Consider the following continuation of the above example.

Example 2.4. Let the joint density function of $(\boldsymbol{X}, \boldsymbol{Y})$ be given as (same as in Example 2.3)

$$f_{\boldsymbol{XY}}(x, y) = \begin{cases} 6\exp(-x-2y), & 0 \leq x < y; \\ 0, & \text{otherwise.} \end{cases}$$

Compute the marginal density functions $f_{\boldsymbol{X}}(x)$ and $f_{\boldsymbol{Y}}(y)$, and compare them to the conditional density functions. What can you say about their differences?

Solution:

■

Given the fact that the conditional distribution represents the "refreshed" belief on the behavior of the random variable $\boldsymbol{X}$, we can define the "refreshed expectation" based on this "refreshed belief". This is formalized as the conditional expectation.

Definition 2.4. The conditional expectation of $\boldsymbol{X}$, conditioned on $\{\boldsymbol{Y}=y\}$, is given by

$$\mu_{\boldsymbol{X}|y}=\mathbb{E}_{\boldsymbol{X}|y}\left[\boldsymbol{X}\right]=\sum_{x\in\mathcal{X}} f_{\boldsymbol{X}|y}(x)x,\ \text{ if } \boldsymbol{X} \text{ is discrete and taking values in } \mathcal{X},$$

or

$$\mu_{\boldsymbol{X}|y}=\mathbb{E}_{\boldsymbol{X}|y}\left[\boldsymbol{X}\right]=\int_{-\infty}^{\infty} x f_{\boldsymbol{X}|y}(x)\mathrm{d}x,\ \text{ if } \boldsymbol{X} \text{ is continuous.}$$

Similarly, the conditional variance is defined by

$$\sigma^2_{\boldsymbol{X}|y}\left[\boldsymbol{X}\right]=\sum_{x\in\mathcal{X}} f_{\boldsymbol{X}|y}(x)(x-\mu_{\boldsymbol{X}|y})^2,\ \text{ if } \boldsymbol{X} \text{ is discrete and taking values in } \mathcal{X},$$

or

$$\sigma^2_{\boldsymbol{X}|y}=\int_{-\infty}^{\infty}(x-\mu_{\boldsymbol{X}|y})^2 f_{\boldsymbol{X}|y}(x)\mathrm{d}x,\ \text{ if } \boldsymbol{X} \text{ is continuous.}$$

Despite the above seemingly complex notations, computing the conditional expectation/variance is really the same as computing the un-conditional ones, after we have computed the conditional mass/density function.

Example 2.5. Let the joint density function of $(\boldsymbol{X},\boldsymbol{Y})$ be given as (same as in Example 2.3)

$$f_{\boldsymbol{XY}}(x,y)=\begin{cases}6\exp(-x-2y),\ 0\le x<y;\\ 0,\ \text{otherwise.}\end{cases}$$

Compute the conditional expectation and the variance of $\boldsymbol{X}$, given that $\boldsymbol{Y}=1$.

Solution:

■

2.1 Independence

We have seen examples showing that in general, the marginal density/mass function $f_{\boldsymbol{X}}$ is different from the conditional density/mass function $f_{\boldsymbol{X}|y}$. This is in general due to the following observation:

There can exist coupling between $\boldsymbol{X}$ and $\boldsymbol{Y}$, so knowing the value of $\boldsymbol{Y}$ changes the probabilistic behavior of $\boldsymbol{X}$.

Consider a simple example.

Example 2.6. Suppose

$$\boldsymbol{X} \sim \text{Bernoulli}(p), \;\; \boldsymbol{Y} = \boldsymbol{X}.$$

It is clear that both $\boldsymbol{X}$ and $\boldsymbol{Y}$ follow the Bernoulli distribution with parameter $p \in (0, 1)$. Now conditioned on $\{\boldsymbol{Y} = 1\}$, what is the distribution of $\boldsymbol{X}$?

Solution:

■

The above example shows a simple example of "coupling" between two random variables.

Now how do we define the notion that is the opposite of "coupling" – meaning that knowing the value of one random variable has given us no information on the value of the other random variable? This is formalized as independence.

Definition 2.5. Two random variables $(\boldsymbol{X}, \boldsymbol{Y})$ (regardless of being continuous or discrete) are independent if and only if

$$f_{\boldsymbol{X}\boldsymbol{Y}}(x, y) = f_{\boldsymbol{X}}(x) \cdot f_{\boldsymbol{Y}}(y).$$

We also have the following equivalent definitions of independence.

Proposition 2.2. Two random variables $(\boldsymbol{X}, \boldsymbol{Y})$ (regardless of being continuous or discrete) are independent if and only if

1. We have $f_{\boldsymbol{X}|y}(x) = f_{\boldsymbol{X}}(x)$ whenever y satisfies $f_{\boldsymbol{Y}}(y) > 0$.
2. We have $f_{\boldsymbol{Y}|x}(y) = f_{\boldsymbol{Y}}(y)$ whenever x satisfies $f_{\boldsymbol{X}}(x) > 0$.
3. We have $\mathbb{P}(\boldsymbol{X} \in A, \boldsymbol{Y} \in B) = \mathbb{P}(\boldsymbol{X} \in A)\mathbb{P}(\boldsymbol{Y} \in B)$ for any A in tha range of $\boldsymbol{X}$ and B in the range of $\boldsymbol{Y}$.

Since the above proposition is so important, we now proceed to establish the above claims at least for discrete random variables.

Proof.

□

The last claim in Proposition 2.2 also provides us with a simple way to compute the probability involving two random variables. This is demonstrated in the following example.

Example 2.7. Let the joint density function of $(\boldsymbol{X}, \boldsymbol{Y})$ be given as (same as in Example 2.3)

$$f_{\boldsymbol{XY}}(x, y) = \begin{cases} 6\exp(-x-2y), \ 0 \le x < y; \\ 0, \ \text{otherwise.} \end{cases}$$

Questions: are $\boldsymbol{X}$ and $\boldsymbol{Y}$ independent? Compute the following probability:

$$\mathbb{P}(0 \le \boldsymbol{X} \le 1, \ 0 \le \boldsymbol{Y} \le 1) = ?$$

Consider another setting. Let the joint density function of $(\boldsymbol{X}, \boldsymbol{Y})$ be given as (same as in Example 2.3)

$$f_{\boldsymbol{XY}}(x, y) = \begin{cases} 2\exp(-x-2y), \ 0 \le x, 0 \le y; \\ 0, \ \text{otherwise.} \end{cases}$$

Questions: are $\boldsymbol{X}$ and $\boldsymbol{Y}$ independent? Compute the following probability:

$$\mathbb{P}(0 \leq \boldsymbol{X} \leq 1,\ 0 \leq \boldsymbol{Y} \leq 1) =?$$

Solution:

$\blacksquare$

Let us consider the following example that can generalize the notion of independence.

Proposition 2.3. Let $(\boldsymbol{X}, \boldsymbol{Y})$ be independent. For any function $h(\cdot)$ of $\boldsymbol{X}$, we have $(h(\boldsymbol{X}), \boldsymbol{Y})$ are also independent.

Despite the claim being very intuitive, its reasoning is an excellent exercise for our understanding of independence.

Proof. The reasoning is fairly simple for discrete random variables.

□

Why is Proposition 2.3 useful? Let us try to apply it in the following example.

Example 2.8. Let the joint density function of $(\boldsymbol{X}, \boldsymbol{Y})$ be given as (same as in Example 2.3)

$$f_{\boldsymbol{XY}}(x, y) = \begin{cases} 2\exp(-x-2y), & 0 \leq x, 0 \leq y; \\ 0, & \text{otherwise.} \end{cases}$$

Compute the following probability:

$$\mathbb{P}(0 \leq \boldsymbol{X}^3 \leq 8,\ 0 \leq \boldsymbol{Y} \leq 1) = ?$$

Solution:

■

3 Covariance and Correlation

Our motivating questions is:

If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are not independent with each other? They how much do they interfere with each other?

This can be roughly summarized by the correlation of $\boldsymbol{X}$ and $\boldsymbol{Y}$. But before that, let us define the expectation of functions of $(\boldsymbol{X}, \boldsymbol{Y})$.

Definition 3.1. For any function $h(x, y)$, the expecation of $h(\boldsymbol{X}, \boldsymbol{X})$ is defined as

$$\mathbb{E}\left[h(\boldsymbol{X}, \boldsymbol{Y})\right] = \begin{cases} \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} h(x, y) f_{\boldsymbol{XY}}(x, y), & \text{if } (\boldsymbol{X}, \boldsymbol{Y}) \text{ are discrete,} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{\boldsymbol{XY}}(x, y) \mathrm{d}y \mathrm{d}x, & \text{if } (\boldsymbol{X}, \boldsymbol{Y}) \text{ are continuous.} \end{cases}$$

It should be noted that the above definition covers the expectation and the variance of $\boldsymbol{X}$ (or $\boldsymbol{Y}$) as special cases.

Definition 3.2. The covariance between $\boldsymbol{X}$ and $\boldsymbol{Y}$ is defined as

$$\sigma_{\boldsymbol{XY}} = \mathbb{E}\left[(\boldsymbol{X} - \mu_{\boldsymbol{X}})(\boldsymbol{Y} - \mu_{\boldsymbol{Y}})\right] = \mathbb{E}\left[\boldsymbol{XY}\right] - \mu_{\boldsymbol{X}}\mu_{\boldsymbol{Y}}.$$

Exercise 3.1. Why does the following hold in the definition of covariance?

$$\mathbb{E}\left[(\boldsymbol{X}-\mu_{\boldsymbol{X}})(\boldsymbol{Y}-\mu_{\boldsymbol{Y}})\right] = \mathbb{E}\left[\boldsymbol{X}\boldsymbol{Y}\right] - \mu_{\boldsymbol{X}}\mu_{\boldsymbol{Y}}?$$

Solution:

$\blacksquare$

Below are some basic properties of covariance.

Proposition 3.1. We have

$$\sigma_{\boldsymbol{X}\boldsymbol{Y}} = \sigma_{\boldsymbol{Y}\boldsymbol{X}}.$$

In addition, if $\boldsymbol{Z} = c \cdot \boldsymbol{X}$ for some constant c, then

$$\sigma_{\boldsymbol{Z}\boldsymbol{Y}} = c \cdot \sigma_{\boldsymbol{X}\boldsymbol{Y}}.$$

Proof.

$\square$

To see how covariance reflects the "coupling" between $\boldsymbol{X}$ and $\boldsymbol{Y}$, let us consider the following example

Example 3.1. Consider $\boldsymbol{X} \sim N(0,1)$ be a standard normal random variable. Let $\boldsymbol{Y} = \boldsymbol{X}$, and let $\boldsymbol{Z} = -\boldsymbol{X}$, Compute the covariance $\sigma_{\boldsymbol{X}\boldsymbol{Y}}$ and $\sigma_{\boldsymbol{X}\boldsymbol{Z}}$.

Moreover, define $U = (\boldsymbol{Y}+\boldsymbol{Z})/2$, compute the covariance $\sigma_{\boldsymbol{X}\boldsymbol{U}}$.

Question: how would you describe the coupling/relation between $\boldsymbol{X}$ and $(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{U})$? Does the covariance reflect this coupling?

Solution:

■

On the other hand, covariance itself is not the right notion when describing the strength of "coupling" between two random variables. Consider the following exercise.

Exercise 3.2. Fixing $\boldsymbol{X}$, if for some $\boldsymbol{Y}, \boldsymbol{Z}$, we have $\sigma_{\boldsymbol{XY}} > \sigma_{\boldsymbol{XZ}}$, then does this mean that $\boldsymbol{Y}$ is more correlated with $\boldsymbol{X}$ compared to $\boldsymbol{Z}$ being more "coupled" with $\boldsymbol{X}$?

Solution: Hint: consider $\boldsymbol{Y} = 10^9 \cdot \boldsymbol{X}$, $\boldsymbol{Z} = 0.5 \cdot \boldsymbol{X}$.

■

The above exercise shows that we have to take into account of the variance of $(\boldsymbol{X}, \boldsymbol{Y})$ when measuring the "coupling" strength between $\boldsymbol{X}$ and $\boldsymbol{Y}$. Consequently, this motivates the notion of correlation.

Definition 3.3. The correlation between $\boldsymbol{X}$ and $\boldsymbol{Y}$ is defined as

$$\rho_{\boldsymbol{XY}} = \frac{\sigma_{\boldsymbol{XY}}}{\sigma_{\boldsymbol{X}}\sigma_{\boldsymbol{Y}}}.$$

As the first step, we make note of the following simple fact.

Example 3.2. The covariance and the correlation between $\boldsymbol{X}$ and itself is

$$\sigma_{\boldsymbol{XX}} = \sigma_{\boldsymbol{X}}^2,\ \rho_{\boldsymbol{XX}} = 1.$$

Solution: Hint: by definition.

■

From the above example, at least we know that for the simplest case $\boldsymbol{Y} = \boldsymbol{Y}$, the notion of covariance and correlation is very intuitive.

The next example shows that correlation is always constrained between -1 and $+1$, for any $\boldsymbol{X}$ and $\boldsymbol{Y}$.

Example 3.3. Show that for any $(\boldsymbol{X}, \boldsymbol{Y})$, we have

$$-1 \leq \rho_{\boldsymbol{XY}} \leq 1.$$

Proof.

□

Knowing correlation is bounded between $[-1, +1]$ is a good thing, since we know that we can no longer manipulate it to arbitrarily large values as we did for the covariance. Moreover, the next proposition shows that corelation is a very natural metric for measuring the degree of independence (meaning that correlation is zero whenever we have independence).

Proposition 3.2. If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent, then

$$\sigma_{\boldsymbol{XY}} = \rho_{\boldsymbol{XY}} = 0.$$

Proof.

□

What happens if $\rho_{\boldsymbol{XY}}$ takes the value of $+1$ or -1? What can we tell about the relation between $\boldsymbol{X}$ and $\boldsymbol{Y}$?

It turns out that the correlation is a VERY useful indicator for the potential linear relationship between $\boldsymbol{X}$ and $\boldsymbol{Y}$;

Proposition 3.3. We have $\rho_{\boldsymbol{XY}} = 1$ if and only if $\boldsymbol{Y} = c \cdot \boldsymbol{X} + b$ for some $c > 0$ and $b \in \mathbb{R}$.

We have $\rho_{\boldsymbol{XY}} = -1$ if and only if $\boldsymbol{Y} = -c \cdot \boldsymbol{X} + b$ for some $c > 0$ and $b \in \mathbb{R}$.

Proof.

□

Finally, let us note that the reverse direction of Proposition 3.2 does not hold unless for very special case.

Proposition 3.4. In general,

$$\sigma_{\boldsymbol{XY}} = \rho_{\boldsymbol{XY}} = 0 \not\Rightarrow \ \boldsymbol{X} \text{ and } \boldsymbol{Y} \text{ are independent.} \tag{3.1}$$

But if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are normal random variables. Then

$$\sigma_{\boldsymbol{XY}} = \rho_{\boldsymbol{XY}} = 0 \Rightarrow \ \boldsymbol{X} \text{ and } \boldsymbol{Y} \text{ are independent.} \tag{3.2}$$

Proof. The second claim (3.2) requires some advanced tools. Let us show only claim (3.1).

□

4 Multiple Random Variables and Their Linear Functions

This section focus on deriving the expectation and the variance for linear functions applied to multiple random variables.

First, let us define the joint mass or density functions for multiple random variables.

4.1 Joint Density/Mass Function, Independence

Definition 4.1 (Joint Probability Mass Function for Multiple Random Variables)**.** Suppose $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are discrete random variables, their joint mass function is a function satisfying

1. $f_{\boldsymbol{X}_1 \ldots \boldsymbol{X}_n}(x_1, \ldots, x_n) \geq 0, \; \forall x_1, \ldots, x_n$.
2. $\sum_{x_1, \ldots, x_n} f_{\boldsymbol{X}_1 \ldots \boldsymbol{X}_n}(x_1, \ldots, x_n) = 1$.
3. $f_{\boldsymbol{X}_1 \ldots \boldsymbol{X}_n}(x_1, \ldots, x_n) = \mathbb{P}(\boldsymbol{X}_1 = x_1, \ldots, \boldsymbol{X}_n = x_n)$.

Definition 4.2 (Joint Probability Density Function for Multiple Random Variables)**.** Suppose $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are continuous random variables, their joint density function is a function satisfying

1. $f_{\boldsymbol{X}_1 \ldots \boldsymbol{X}_n}(x_1, \ldots, x_n) \geq 0, \; \forall x_1, \ldots, x_n$.
2. $\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\boldsymbol{X}_1 \ldots \boldsymbol{X}_n}(x_1, \ldots, x_n) \mathrm{d}x_1 \cdots \mathrm{d}x_n = 1$.
3. For any B that is a subset of $\mathbb{R}^n$ (the n-dimensional space), we have
$$\int_{(x_1, \ldots, x_n) \in B} f(x_1, \ldots, x_n) \mathrm{d}x_1 \cdots \mathrm{d}x_n = \mathbb{P}((\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) \in B).$$

Accordingly, the marginal mass/density function is defined in exactly the same way as before.

Definition 4.3 (Marginal Probability Mass/Density Function)**.** Suppose $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are discrete random variables, the marginal mass function of $\boldsymbol{X}_i$ is given as
$$f_{\boldsymbol{X}_i}(x_i) = \sum_{x_1} \cdots \sum_{x_{i-1}} \sum_{x_{i+1}} \cdots \sum_{x_n} f_{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n}(x_1, \ldots, x_n).$$

Suppose $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are continuous random variables, the marginal density function of $\boldsymbol{X}_i$ is given as

$$f_{\boldsymbol{X}_i}(x_i) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n}(x_1, \ldots, x_n) \mathrm{d}x_1 \cdots \mathrm{d}x_{i-1} \mathrm{d}x_{i+1} \cdots \mathrm{d}x_n.$$

With the above definitions in place, we can talk about the independence of multiple random variables.

Definition 4.4. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be random variables (they can be jointly discrete or continuous). Then $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are independent if and only if

$$f_{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n}(x_1, \ldots, x_n) = f_{\boldsymbol{X}_1}(x_1) \cdots f_{\boldsymbol{X}_n}(x_n), \ \forall (x_1, \ldots, x_n).$$

The above definition might seem not intuitive. The following equivalent definition shows the true power of independence.

Proposition 4.1. Suppose $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are independent (they can be jointly discrete or continuous), if and only if for any $A_1, \ldots, A_n$ that are in the range of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$:

$$\mathbb{P}(\boldsymbol{X}_1 \in A_1, \ldots, \boldsymbol{X}_n \in A_n) = \mathbb{P}(\boldsymbol{X}_1 \in A_1) \cdots \mathbb{P}(\boldsymbol{X}_n \in A_n).$$

Proof. Hint: we can just consider discrete random variables to gain sufficient intuition for understanding this proposition. Can you re-use our argument for showing Proposition 2.2?

□

Given the above proposition, it is clear that for independent random variables, the joint probability can be simplified as the product of multiple marginal probabilities.

Exercise 4.1. Suppose $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are independent, then

$$\text{Are } \boldsymbol{X}_1 \text{ and } \boldsymbol{X}_2 \text{ independent?}$$

Solution: Hint: can we first recall the equivalent definition (Proposition 2.2 - 3) of independence for two random variables?

■

4.2 Expectation and Variance of Linear Functions

We can now begin to talk about the expectation and the variance of linear function applying to multiple random variables.

Proposition 4.2. Given constants $c_1, \ldots c_n$, and let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be random variables (they can be either continuous or discrete), define

$$\boldsymbol{Y} = c_1 \boldsymbol{X}_1 + \cdots + c_n \boldsymbol{X}_n.$$

Clearly, $\boldsymbol{Y}$ is the result of linear function $h(x_1, \ldots, x_n) = c_1 x_1 + c_n x_n$ applied to $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. We have

$$\mathbb{E}\left[\boldsymbol{Y}\right] = \mathbb{E}\left[c_1 \boldsymbol{X}_1 + \cdots + c_n \boldsymbol{X}_n\right] = c_1 \mathbb{E}\left[\boldsymbol{X}_1\right] + \cdots + c_n \mathbb{E}\left[\boldsymbol{X}_n\right].$$

A very important feature of the above property is that we do not care about the relationship between random variables $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$!

Example 4.1. Suppose $\boldsymbol{X}$ follows $N(0, 1)$, $\boldsymbol{Y}$ follows $N(0, 1)$, and $\rho_{\boldsymbol{XY}}(\boldsymbol{X}, \boldsymbol{Y}) = 0.5$. Question:

$$\mathbb{E}\left[2\boldsymbol{X} + 3\boldsymbol{Y}\right] = ?$$

Solution:

■

Now the variance of linear functions applied to multiple random variables seems more complicated. We will also see that computing such a quantity involves the covariance we previously defined .

Proposition 4.3. Given constants $c_1, \ldots c_n$, and let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be random variables (they can be either continuous or discrete), define

$$\boldsymbol{Y} = c_1\boldsymbol{X}_1 + \cdots + c_n\boldsymbol{X}_n.$$

Clearly, $\boldsymbol{Y}$ is the result of linear function $h(x_1, \ldots, x_n) = c_1x_1 + c_nx_n$ applied to $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. We have

$$\begin{aligned} V(\boldsymbol{Y}) = V(c_1\boldsymbol{X}_1 + \cdots + c_n\boldsymbol{X}_n) &= c_1^2V(\boldsymbol{X}_1) + \cdots c_n^2V(\boldsymbol{X}_n) + 2\sum_{i<j} c_ic_j\sigma_{\boldsymbol{X}_i\boldsymbol{X}_j} \\ &= c_1^2V(\boldsymbol{X}_1) + \cdots c_n^2V(\boldsymbol{X}_n) + \sum_{i\neq j} c_ic_j\sigma_{\boldsymbol{X}_i\boldsymbol{X}_j} \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} c_ic_j\sigma_{\boldsymbol{X}_i\boldsymbol{X}_j}. \end{aligned}$$

If in addition, that $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are independent, then

$$V(\boldsymbol{Y}) = V(c_1\boldsymbol{X}_1 + \cdots + c_n\boldsymbol{X}_n) = c_1^2V(\boldsymbol{X}_1) + \cdots c_n^2V(\boldsymbol{X}_n).$$

Before showing the reasoning behind the above proposition. Let us first apply it to the following example.

Example 4.2. Suppose $\boldsymbol{X}$ follows $N(0,1)$, $\boldsymbol{Y}$ follows $N(0,1)$, and $\rho_{\boldsymbol{XY}}(\boldsymbol{X}, \boldsymbol{Y}) = 0.5$. Question:

$$V(2\boldsymbol{X} + 3\boldsymbol{Y}) = ?$$

Solution:

■

Proof of Proposition 4.3. The proof is indeed an excellent exercise by using the definition of variance, covariance, and the expectation of linear functions applied to multiple random variables (Proposition 4.2).

□

Of course, we will be paying additional attention to some special linear functions. It turns out this linear function possesses some very useful properties, and we will repeatedly encounter them during our discussions on statistics.

Proposition 4.4 (Mean and Variance of Averages for Independent Random Variables)**.** Suppose $\overline{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i$, where $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are independent, with

$$\mathbb{E}\left[\boldsymbol{X}_i\right] = \mu,\ V(\boldsymbol{X}_i) = \sigma^2,\ \forall i = 1, \ldots n.$$

Then we have

$$\mathbb{E}\left[\overline{\boldsymbol{X}}\right] = \mu,\ V(\overline{\boldsymbol{X}}) = \frac{\sigma^2}{n}.$$

To see the application of the above proposition. Let us consider the following example.

Example 4.3. Suppose the length of a box is 10 mm. We take 100 independent measurements, with each of them following $N(10, 1)$. That is, the error of each measurement is $N(0, 1)$.

What would be the strategy that we can use to obtain an estimator of the box length that is of high accuracy?

Solution:

■

The last property we introduce in this chapter is the so-called Reproductive Property of the Normal Distribution.

Proposition 4.5. Suppose each $\boldsymbol{X}_i$ follows $N(\mu_i, \sigma_i^2)$, for $i = 1, \ldots n$. For any constant $c_1, \ldots, c_n$, define

$$\boldsymbol{Y} = c_1\boldsymbol{X}_1 + \cdots + c_n\boldsymbol{X}_n.$$

Then $\boldsymbol{Y}$ also follows a normal distribution. More precisely, $\boldsymbol{Y}$ follows

$$N(\sum_{i=1}^{n} c_i\mu_i, \ \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \sigma_{\boldsymbol{X}_i\boldsymbol{X}_j}).$$

If in particular, that $\boldsymbol{X}_1, \ldots \boldsymbol{X}_n$ are independent, then $\boldsymbol{Y}$ follows

$$N(\sum_{i=1}^{n} c_i\mu_i, \ \sum_{i=1}^{n} c_i^2\sigma_i^2).$$

Example 4.4. Suppose $\boldsymbol{X}$ follows $N(0, 1)$, $\boldsymbol{Y}$ follows $N(0, 1)$, and $\rho_{\boldsymbol{XY}}(\boldsymbol{X}, \boldsymbol{Y}) = 0.5$. Question:

What is the distribution of $\boldsymbol{Z} = 2\boldsymbol{X} + 3\boldsymbol{Y}$?

Solution:

■

Proposition 4.6. If $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are n discrete random variables, and let $f_{\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n}(x_1, \ldots, x_n)$ be their joint probability mass function, then for any function $h(x_1, \ldots, x_n)$, we have

$$\mathbb{E}\left[h(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)\right] = \sum_{x_1}\sum_{x_2}\cdots\sum_{x_n} h(x_1, \ldots, x_n) f_{\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n}(x_1, \ldots, x_n).$$

If $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are n continuous random variables, and let $f_{\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n}(x_1, \ldots, x_n)$ be their joint probability density function, then for any function $h(x_1, \ldots, x_n)$, we have

$$\mathbb{E}\left[h(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)\right] = \int\int\cdots\int h(x_1, \ldots, x_n) f_{\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n}(x_1, \ldots, x_n) \mathrm{d}x_1 \cdots \mathrm{d}x_n.$$

Example 4.5. If we have continuous random variable $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$, then

1. Let $h(x, y, z) = x$, then

$$\mathbb{E}\left[h(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})\right] = \mathbb{E}\left[\boldsymbol{X}\right] = \int\int\int x f_{XYZ}(x, y, z) \mathrm{d}x\mathrm{d}y\mathrm{d}z$$

2. Let $h(x, y, z) = xy$, then

$$\mathbb{E}\left[h(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})\right] = \mathbb{E}\left[\boldsymbol{X}\boldsymbol{Y}\right] = \int\int\int xy f_{XYZ}(x, y, z) \mathrm{d}x\mathrm{d}y\mathrm{d}z$$

ISyE 3770: Point Estimation

This chapter begins our discussion of statistics this semester. Indeed, statistics can be roughly categorized into two topics. The first is called parameter estimation, and the other is called hypothesis testing. This chapter will focus on parameter estimation.

1 Concepts of Point Estimation

We start with the following example.

Example 1.1. Suppose we want to estimate the average height among people living in the US.

Suppose we have randomly selected $n = 100$ people (with replacement) among all the people in the US. Let us denote $\boldsymbol{X}_i$ as the i-th selected person's height. It is clear that $\boldsymbol{X}_i$ is random.

A point estimator for the true average:

$$\boldsymbol{Y} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i$$

The above $\boldsymbol{Y}$ that utilizes the collected samples $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ is called a point estimator of the true average height of US population.

An important observation: Note that $\boldsymbol{Y}$ is a function of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$! That is, define function

$$h(x_1, \ldots, x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

we have

$$\boldsymbol{Y} = h(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n).$$

Now since $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are random variables, this implies that our point estimator $\boldsymbol{Y}$ is also a random variable!

Definition 1.1 (Point Estimator)**.** Suppose we want to estimate some quantity p related to the population (generally unknown since the population is typically prohibitively large for us to enumerate). Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be the collected samples.

In order to estimate p, suppose we have specified a function $h(x_1, \ldots, x_n)$ and define

$$\Theta = h(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n).$$

Then $\Theta(\boldsymbol{X}_1, \ldots \boldsymbol{X}_n)$ is called a point estimator of p.

It should be noted that Θ is a random variable itself! As it is a function of random variables $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$.

There is one point in the above example that we overlooked but is very important to highlight. That is, every person's height that is selected has the same distribution as others!

Definition 1.2 (Independent and Identically Distributed Samples)**.** We say $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are independently and identically distributed (i.i.d. as the shorthand) samples if

1. Every $\boldsymbol{X}_i$ has the same distribution.
2. Different $\boldsymbol{X}_i$ are independent.

Why do we care about the above two properties for drawing random samples?

Example 1.2. Consider the following scenarios in Example 1.1.

1. The first 90 persons are sampled from the east, and later 10 persons are sampled from the west coast.
2. For some technical issue, starting from the 3nd person, our random selector has a bug and we repeatedly sample the 2st person over and over again.

Can we identify which condition of I.I.D. is broken in each of the above scenarios?

Solution:

■

Lastly, the point estimator can also be viewed as a special case of the notion called statistics.

Definition 1.3 (Statistics)**.** Given random samples $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, and a function $h(x_1, \ldots, x_n)$, the random variable $h(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ is called a statistics.

Remark 1.1. It should be noted that statistics can also be viewed as a point estimator – except that it does not necessarily need to estimate anything.

For instance, in Example 1.1, one can define

$$\boldsymbol{Y} = 2\boldsymbol{X}_1^{34428735} - 5\boldsymbol{X}_{100},$$

then $\boldsymbol{Y}$ is a function of $\boldsymbol{X}_1, \boldsymbol{X}_{100}$ (and consequently a function of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{100}$), hence $\boldsymbol{Y}$ is a valid statistics.

But, $\boldsymbol{Y}$ does not really try to estimate anything.

Since the point estimator itself is a statistic – a function of random samples - which also shows that point estimator (statistics) is a random variable.

It is then very natural to consider the distribution of a statistic.

Definition 1.4 (Sampling Distribution)**.** The probability distribution of a statistic is called the sampling distribution of this statistic.

At this point, it might appear that there is little motivation for us to suddenly dive into the distribution of the point estimator. We will illustrate this through an example.

Example 1.3 (Importance of Sampling Distribution)**.** Suppose in Example 1.1, the averaged height in US population is p meters, and

$$\boldsymbol{X}_i \sim N(p, 0.1),$$

that is, each randomly sampled person from US population has a height p meters and varaince 0.1 meters^2. Here p is unknown.

In addition, take $n = 1000$

$$\boldsymbol{Y} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i$$

is a point estimator (statistic).

Now given our discussion in the last Chapter, we then know that

$$\boldsymbol{Y} \sim N(p, 10^{-4}). \tag{1.1}$$

This will imply the following:

$$\begin{aligned}
\mathbb{P}(\boldsymbol{Y} - p \leq -3 \cdot 10^{-2} \text{ or } \boldsymbol{Y} - p \geq 3 \cdot 10^{-2}) &= \mathbb{P}(\frac{\boldsymbol{Y} - p}{10^{-2}} \leq \frac{-3 \cdot 10^{-2}}{10^{-2}}) + \mathbb{P}(\frac{\boldsymbol{Y} - p}{10^{-2}} \geq \frac{3 \cdot 10^{-2}}{10^{-2}}) \\
&= \Phi(\frac{-3 \cdot 10^{-2}}{10^{-2}}) + 1 - \Phi(\frac{3 \cdot 10^{-2}}{10^{-2}}) \\
&= \Phi(-3) + 1 - \Phi(3) \approx 0.02.
\end{aligned}$$

Equivalently, we know that

$$\mathbb{P}(p - 0.03 \leq \boldsymbol{Y} \leq p + 0.03) = \mathbb{P}(\boldsymbol{Y} - 0.03 \leq p \leq \boldsymbol{Y} + 0.03) = 0.02.$$

In other words, we claim that

Suppose we have sampled 1000 persons and computed the average height among them, denoted as $\boldsymbol{Y}$. Then with a probability at least 0.98, the (unknown) true average height among US population is guaranteed to fall in the small neighborhood $[\boldsymbol{Y} - 0.03, \boldsymbol{Y} + 0.03]$!

The above example clearly demonstrates why it is important to dive into the distribution of the point estimator – typically the unknown quantity of interest (e.g., the average height of population) will appear in the sampling distribution (e.g. (1.1)). With the sampling distribution known to us, we can then make a probabilistic statement on how close our point estimator is compared to the quantity of interest.

Now (1.1) in Example 1.3 is pretty straightforward to obtain. This should be note taken for granted! In fact, obtaining sampling distribution for statistics $\boldsymbol{Y} = h(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ is often highly non-trivial.

Consider the following example.

Example 1.4. Suppose we have drawn two people from the population, each person's height is a random variable $N(1.8, 0.1)$. Can we determine what is the distribution of $\boldsymbol{Y} = \boldsymbol{X}_1/\boldsymbol{X}_2$? Here X_i denotes the i-th person's height.

What about $\boldsymbol{Y} = \boldsymbol{X}_1^2/\boldsymbol{X}_2$?

Solution:

■

The above example shows that obtaining the sampling distribution for general statistics can be tricky or sometimes impossible. Fortunately, in common scenarios, the statistics is not a potentially unfriendly function of collected samples. Rather, it is a simple linear combination of the collected samples. In this case, we can apply the following well-known central limit theorem.

Theorem 1.1 (Central Limit Theorem)**.** *Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be I.I.D. samples drawn from a distribution with mean μ and variance σ^2. Note that (μ, σ^2) is defined by the distribution, not computed by these n samples.*

Let $\overline{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i$ be the sample average (mean), then

$$\boldsymbol{Z} = \frac{\boldsymbol{X} - \mu}{\sigma/\sqrt{n}}$$

follows approximately standard normal distribution $N(0, 1)$. The approximation becomes exact as $n \to \infty$, meaning the limiting distribution of $\boldsymbol{Z}$, as $n \to \infty$, is exactly $N(0, 1)$.

It is immediate to obtain the following corollary that can sometimes be useful.

Corollary 1.1. Under the same setup as Theorem 1.1. Then $\overline{\boldsymbol{X}}$ approximately follows a distribution of

$$N(\mu, \frac{\sigma^2}{n}).$$

Consequently, $\overline{\boldsymbol{X}}$ fluctuates around the true mean μ, and the error $\overline{\boldsymbol{X}} - \mu$ follows a distribution $N(0, \frac{\sigma^2}{n})$.

Clearly, as n goes to infinity, the error shrinks to 0 as the variance of the normal distribution shrinks to 0 at a rate of $O(1/n)$.

Proof.

□

Let us see some applications of central limit theorem.

Example 1.5 (Averages of Uniform R.V.s)**.** Suppose each $\boldsymbol{X}_i$ follows a uniform distribution over $[4, 6]$. We can readily compute that

$$\mu = 5, \sigma^2 = 1/3.$$

Consequently, applying Central Limit Theorem shows for $n = 40$:

$$\overline{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \text{ approximately follows distribution } N(5, \frac{1}{120}).$$

Can you draw a graph for the P.D.F. of $\overline{\boldsymbol{X}}$ and visualize the above phenomenon?

Solution:

■

2 Some Concepts of Point Estimation

Let us we have a distribution which contains an unknown parameter θ. We can imagine the scenario where we know the height of US individuals follows $N(\theta, 0.01)$, but with an unknown θ.

In the previous section we introduce the concept of an estimator $(\widehat{\theta}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n))$, which is a function of n samples collected by sampling from this distribution, and aims at estimating the unknown θ.

When the context is clear, we will write $\widehat{\theta}$ in short for $\widehat{\theta}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$.

Definition 2.1 (Bias of an estimator)**.** The bias of the estimator $\widehat{\theta}$ is defined as

$$\text{Bias}(\widehat{\theta}) = \mathbb{E}\left[\widehat{\theta}\right] - \theta.$$

Exercise 2.1. Why do we take expectation in the definition of the bias? What is random there?

Solution:

■

Example 2.1. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be the I.I.D. sample from distribution with mean μ, and variance σ^2. Let us consider the sample mean, and sample variance:

$$\overline{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i,\ S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\boldsymbol{X}_i - \overline{\boldsymbol{X}}\right)^2$$

We claim that sample mean and sample variance has zero bias:

$$\mathbb{E}\left[\overline{\boldsymbol{X}}\right] = \mu,\ \mathbb{E}\left[S^2\right] = \sigma^2.$$

Proof. This is an excellent exercise to refresh our knowledge on the normal distribution. □

Solution:

■

Definition 2.2 (Variance of an estimator)**.** The variance of an estimator $\widehat{\theta}$ is simply

$$V(\widehat{\theta}).$$

Exercise 2.2. Why do we care about the variance of an estimator $\widehat{\theta}$?

Can we think about a scenario where we can use the variance to judge the quality of two unbiased estimators?

Hint: US population example – compare the cases where we have 10 people and 1000 people's height.

Solution:

■

There is an important class of estimators that is worth of our dedicated attention.

Definition 2.3 (Minimum variance unbiased estimator)**.** Fixing the number of samples n, among all possible choices of estimators $\widehat{\theta}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$, the ones that are both unbiased and achieve the minimal variance among the unbiased estimators are called minimum variance unbiased estimator (MVUE).

Finding the MVUE estimator for general distributions are typically difficult. But for normal distributions, again we have a clear and natural answer.

Proposition 2.1. Suppose $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are I.I.D. sampled from normal distribution $N(\mu, \sigma^2)$. Then

$$\overline{\boldsymbol{X}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i$$

is the MVUE estimator for the parameter μ.

Now, Exercise 2.2 may suggest that we only need to care about the variance of an estimator. But there is a big caveat, we only care about the variance if the estimator is unbiased!

Definition 2.4 (Mean-squared error of an estimator)**.** The mean-squared error of an estimator $\widehat{\theta}$ is defined as

$$\mathrm{MSE}(\widehat{\theta}) = \mathbb{E}\left[(\widehat{\theta} - \theta)^2\right].$$

To see how the mean-squared error relates the bias and the variance of an estimator, we now introduce the following **very important** relation that is frequently used in statistical analysis.

Proposition 2.2. We have the bias and variance decomposition of the mean-squared error:

$$\mathrm{MSE}(\widehat{\theta}) = \mathrm{Bias}(\widehat{\theta})^2 + V(\widehat{\theta}).$$

Proof.

□

3 Common Methods of Point Estimation

So far we have mostly talked about given a point estimator, how do we measure its quality. A natural question is then how do we actually obtain a reasonable estimator for some quantity of interest? In this section we will introduce two common approaches for doing so.

3.1 Method of Moments

Definition 3.1 (Moments)**.** Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_2$ be random variables (either discrete or continuous) from the same distribution. For any $k = 1, 2, \ldots$, the k-th population moment is defined as

$$\mathbb{E}\left[\boldsymbol{X}_1^k\right] = \cdots = \mathbb{E}\left[\boldsymbol{X}_n^k\right],$$

and the k-th sample moment is

$$\frac{1}{n}\sum_{i=1} \boldsymbol{X}_i^k.$$

For example, we have seen examples of computing the first and the second moment of normal distributions.

Example 3.1. Let each of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ follow $N(\mu, \sigma^2)$. Then the 1st population moment is μ, and the 2nd population moment is $\mu^2 + \sigma^2$.

Proof.

□

Definition 3.2. Let $\boldsymbol{X}_1, \dots, \boldsymbol{X}_2$ be random variables (either discrete or continuous) from the same distribution. Suppose the distribution has m parameters we have trying to estimate, denoted by $\theta_1, \dots, \theta_m$.

Suppose we can find certain functions $h_1(\cdot), \dots, h_m(\cdot)$ so that

$$\begin{aligned} \text{1-st population moment} &= h_1(\theta_1, \dots, \theta_m), \\ &\cdots \\ &\cdots \\ \text{m-th population moment} &= h_m(\theta_1, \dots, \theta_m). \end{aligned}$$

Then the moment estimators $(\widehat{\theta}_1, \dots, \widehat{\theta}_m)$ is obtained by solving

$$\begin{aligned} h_1(\widehat{\theta}_1, \dots, \widehat{\theta}_m) &= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i = \text{1-st sample moment}, \\ &\cdots \\ &\cdots \\ h_m(\widehat{\theta}_1, \dots, \widehat{\theta}_m) &= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i^m = \text{m-th sample moment} \end{aligned}$$

We now turn to some concrete examples of moment estimators.

Example 3.2 (Normal distributions)**.** Let $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ be sampled from $N(\mu, \sigma^2)$, where both μ and σ^2 are unknown. Show that the moment estimators of μ and σ^2 are

$$\begin{aligned} \widehat{\mu} &= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i = \overline{\boldsymbol{X}}, \\ \widehat{\sigma^2} &= \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{X}_i - \overline{\boldsymbol{X}})^2 = \frac{n-1}{n} S^2, \end{aligned}$$

where $\overline{\boldsymbol{X}}$ is the sample mean, and S^2 is the sample variance.

Proof.

□

Example 3.3 (Exponential distributions)**.** Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be sampled from an exponential distribution with parameter $\lambda > 0$. Then the moment estiamtor of λ is given by

$$\widehat{\lambda} = \frac{1}{\overline{\boldsymbol{X}}}.$$

Proof.

□

3.2 Method of Maximum Likelihood

Compared to the method of moment, the maximum likelihood estimator is more natural. It is also much more popular than the moment method, and we will see plenty of example/exercises involving the maximum likelihood estimators.

Definition 3.3 (Likelihood)**.** Suppose $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are n I.I.D. samples, each following the distribution $f(x;\theta)$ (this can be either mass function or density function), where θ is a parameter.

Then, given $\boldsymbol{X}_1 = x_1, \ldots, \boldsymbol{X}_n = x_n$, the likelihood function of the collected samples is given by

$$L(\theta; x_1, \ldots, x_n) = f(x_1;\theta) \cdot f(x_2;\theta) \cdots f(x_n;\theta).$$

When the context is clear, we will write $L(\theta)$ in short for $L(\theta; x_1, \ldots, x_n)$.

Example 3.4. The θ above can be a single parameter of interest (we are trying to estimate). For instance, when the distribution of $\boldsymbol{X}_i$ is $N(\mu, 1)$, the mean μ is the parameter of interest.

θ can also be multiple parameters. For instance, when the distribution of $\boldsymbol{X}_i$ is $N(\mu, \sigma^2)$, both the mean μ and the variance σ^2 are the parameter of interest.

Remark 3.1. If the samples $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ are discrete random variables, then the likelihood function $L(\theta)$ is simply the joint probability mass function of $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ at $(x_1, \dots, x_n)$.

If the samples $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ are continuous random variables, then the likelihood function $L(\theta)$ is simply the joint probability density function of $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ at $(x_1, \dots, x_n)$.

Definition 3.4 (Maximum likelihood estimator)**.** Given $\boldsymbol{X}_1 = x_1, \dots, \boldsymbol{X}_n = x_n$, and a proper constraint set Θ, the maximum likelihood estimator (MLE) for θ is defined as

$$\widehat{\theta}(x_1, \dots, x_n) = \operatorname*{argmax}_{\theta \in \Theta} L(\theta; x_1, \dots, x_n).$$

Note that the estimator $\widehat{\theta}$ depends on the samples!

Exercise 3.1.

- Why do we need the constraint set Θ?
- Why can't we simply set $\Theta = \mathbb{R}$?
- Can we come up with an example where $\Theta = [0, 1]$?

Now to obtain the maximum likelihood estimator (MLE), we have to solve the following optimization problem:

$$\widehat{\theta} \leftarrow \max_{\theta \in \Theta} L(\theta; x_1, \dots, x_n), \tag{3.1}$$

where the right hand side is a function of θ, and we need to find the optimal $\widehat{\theta}$ that maximizes this function.

Here is a general recipe that typically works for our purpose.

1. Step 1: Take some proper transformation (**we will see why this can be helpful**):

   $$\ell(\theta; x_1, \dots, x_n) = G(L(\theta; x_1, \dots, x_n)),$$

   where $G(y)$ is a **non-decreasing function**. Consequently, the maximizer of $L(\theta; x_1, \dots, x_n)$ will also be a maximizer of $\ell(\theta; x_1, \dots, x_n)$. For many cases, Step 1 is not needed, and we can simply take $\ell(\theta; x_1, \dots, x_n) = L(\theta; x_1, \dots, x_n)$.

2. Let us consider

   $$\max_{\theta \in \Theta} \ell(\theta; x_1, \dots, x_n).$$

   Now take the gradient of $\ell(\theta; x_1, \dots, x_n)$ w.r.t. θ, and

   $$\text{Find } \overline{\theta} \in \Theta \text{ such that } \nabla \ell(\overline{\theta}; x_1, \dots, x_n) = \mathbf{0}, \tag{3.2}$$

   where $\mathbf{0}$ denotes the zero vector (The reason that we use zero vector is due to the fact that θ can be a vector of multiple parameters. In this case the gradient $\nabla \ell(\overline{\theta}; x_1, \dots, x_n)$ is also a vector). It such $\overline{\theta} \in \Theta$ exists, set $\widehat{\theta} = \overline{\theta}$ to be the MLE estimator.

Caveat. It might exists cases that for certain problems, $\overline{\theta}$ satisfying (3.2) does not exists. In addition, $\overline{\theta}$ can be a minimum of $L(\theta; x_1, \ldots, x_n)$ instead of the maximum! In this case, more advanced optimization tools are needed for solving the problem (3.1). However, most of the exercises and examples we will see in this Chapter can be solved by the above recipe.

Now let us see some concrete examples of MLE, and the application of the above procedure for finding them.

Exercise 3.2. Consider a Bernoulli distribution with parameter $p \in [0, 1]$. Given I.I.D. samples $\boldsymbol{X}_1 = x_1, \ldots, \boldsymbol{X}_n = x_n$. Suppose we want to estimate the parameter p.

1. Write out the likelihood function $L(\theta; x_1, \ldots, x_n)$.
2. Identify the constraint set Θ.
3. Construct the MLE estimator of p.

Solution:

■

Exercise 3.3. Consider an exponential distribution with parameter $\lambda > 0$. Given I.I.D. samples $\boldsymbol{X}_1 = x_1, \ldots, \boldsymbol{X}_n = x_n$, suppose we want to estimate the parameter λ.

1. Write out the likelihood function $L(\lambda; x_1, \ldots, x_n)$.
2. Identify the constraint set Θ.

3. Construct the MLE estimator of p.

Solution:

■

The following example aims to estimate two unknown parameters.

Exercise 3.4. Consider a normal distribution with parameters (μ, σ). Given I.I.D. samples $\boldsymbol{X}_1 = x_1, \ldots, \boldsymbol{X}_n = x_n$, suppose we want to estimate the parameters $\theta = (\mu, \sigma)$.

1. Write out the likelihood function $L(\mu, \sigma; x_1, \ldots, x_n)$.

2. Identify the constraint set Θ.

3. Construct the MLE estimator of (μ, σ).

Solution:

■

So far, all the above MLEs are obtained by following the general recipe of taking the gradient to zero. Below, we show an example for which we can not directly apply this general recipe. Yet the optimization problem we are trying to solve for the MLE is still relatively easy to perform.

This example also illustrates the importance that the MLE depends on the samples!.(I believe this trivial fact is worthy of repeated emphasis).

Exercise 3.5. Consider a uniform distribution over $[0, a]$. That is,

$$f(x; a) = \begin{cases} \frac{1}{a}, & 0 \leq x \leq a; \\ 0, & \text{otherwise.} \end{cases}$$

Given I.I.D. samples $\boldsymbol{X}_1 = x_1, \ldots, \boldsymbol{X}_n = x_n$, suppose we want to estimate the parameter a.

1. Write out the likelihood function $L(a; x_1, \ldots, x_n)$.

2. Identify the constraint set Θ.

3. Construct the MLE estimator of a.

Solution:

■

The following proposition can be very useful in certain occasions.

Proposition 3.1. Let $\widehat{\theta}_1, \ldots, \widehat{\theta}_k$ be the MLE of $\theta_1, \ldots, \theta_k$. For any function $h(\cdot, \ldots, \cdot)$ of k arguments. We have that

$$h(\widehat{\theta}_1, \dots, \widehat{\theta}_k) \text{ is the MLE of } h(\theta_1, \dots, \theta_k).$$

The above proposition might be trivial as it conforms to our intuition. Yet its proof is highly non-trivial. To see the power of the above proposition, consider the following.

Exercise 3.6. Consider a normal distribution $N(0, \sigma^2)$ with parameters σ^2. Given I.I.D. samples $\boldsymbol{X}_1 = x_1, \dots, \boldsymbol{X}_n = x_n$, suppose we want to estimate the parameters σ^2.

Approach 1:

1. Write out the likelihood function $L(\sigma; x_1, \dots, x_n)$.
2. Identify the constraint set Θ.
3. Construct the MLE estimator of σ^2.

Approach 2: Apply Proposition 3.1 and Exercise 3.4.

Solution:

■

To conclude our discussions for MLE estimators. We briefly mention some of the nice properties of the MLE estimator, which is also one of the reasons why it is a popular procedure for constructing point estimation.

Proposition 3.2. Under fairly general technical conditions. Let $\widehat{\theta}(\boldsymbol{X}_1, \dots, \boldsymbol{X}_n)$ be the MLE of θ using I.I.D. samples $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$, then

1. $\widehat{\theta}$ is approximately unbiased:

$$\mathbb{E}\left[\widehat{\theta}(\boldsymbol{X}_1, \dots, \boldsymbol{X}_n)\right] - \theta \to 0 \text{ as } n \to \infty.$$

2. The variance of $\widehat{\theta}$ becomes the smallest among all the possible construction of estimators that is asymptotically unbiased.

3. $\widehat{\theta}$ approximately follows a normal distribution.

Note that Proposition 3.2 essentially states that the MLE estimator becomes the minimal variance unbiased estimator when given enough samples.

ISyE 3770: Statistical Interval

In the previous chapter, we have seen how to obtain a point estimator for an unknown parameter that is of our interest. In many scenarios having a point estimator is not enough.

For instance, when training a neural network we typically have a ”test accuracy”, which is the trained network applied to a pre-specified test dataset (CIFAR-10 test data contains 10K data points). However, the true performance of the network is measure by applying the network to classify all the possible images, and there is no way that we can perform this procedure as the number of images is infinite. Consequently, the ”test accuracy” is a point estimate of the true performance of the network.

From the above discussion, suppose we have seen that a freshly trained neural network obtain 98% accuracy on the test dataset. We are happy – as it performs well on the benchmark – a fair chance for publication. But when asking about deploying this network in real life, one becomes less certain. That is, we do not know whether the good performance comes from that this trained neural network happens to be ”good” to the particular 10K data points in the dataset – and performs less well on other data points in real life. To really grasp an idea on how the true performance is, we need a ”confidence interval”, which gives us a range $[l, u]$, and states that the true performance of the network falls into this range with certain (say, 95%) probability.

In this chapter, we will discuss how to construct such confidence intervals (CI).

1 General Concepts of Confidence Interval

Definition 1.1 (Two-sided Confidence Interval)**.** Suppose θ is an unknown quantity of interest. Given I.I.D. samples $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, and $0 < \alpha < 1$, suppose we can construct endpoints $\boldsymbol{L}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n), \boldsymbol{U}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$, such that

$$\mathbb{P}(\boldsymbol{L}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) \leq \theta \leq \boldsymbol{U}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)) \geq 1 - \alpha. \tag{1.1}$$

Then we say that $[\boldsymbol{L}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n), \boldsymbol{U}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)]$ is an $(1-\alpha)$-confidence interval for θ.

We say that $1 - \alpha$ is the confidence coefficient (level), $\boldsymbol{L}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ is the lower confidence limit, and $\boldsymbol{U}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ is the upper confidence limit.

Remark 1.1. Note that the endpoints $\boldsymbol{L}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n), \boldsymbol{U}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ are both random variables as they depend on the random samples! In general, they also depend on the value of confidence level (i.e., $1 - \alpha$).

Exercise 1.1. When defining confidence (1.1) we have a probability: what is the randomness here?

Solution:

■

The following is a general procedure to construct confidence intervals. The primary motivation of introducing the general procedure is the following:

1. In the following subsections all the examples/procedures for confidence intervals are simply a specialization of this general procedure.
2. Many special confidence intervals are so popular (basically all the concrete CIs we are going to learn later), that sometimes it gives people the impression that those CI are the only (or the golden) rules for constructing CIs – this can not be further away from the truth. It is important to know that those CIs are only a fraction of CIs one can use.

Proposition 1.1 (General Method for CI)**.** Given I.I.D. $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ samples and a confidence level α. Suppose we want to construct a $(1-\alpha)$-CI for a parameter θ that is in the distribution (either mass or density function) of $\boldsymbol{X}_i$.

1. Construct/Find a statistics (random variable) $\boldsymbol{Y}$ such that
   (a) $\boldsymbol{Y} = g(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n; \theta)$ for some function g. That is, $\boldsymbol{Y}$ is a function of both samples and the unknown quantity θ.
   (b) The distribution (either mass/density function) is known to us, and it does not depend on θ. Consequently, we can find two points, $l_\alpha \le u_\alpha$, such that
   $$\mathbb{P}(l_\alpha \le \boldsymbol{Y} \le u_\alpha) \ge 1 - \alpha.$$
   One possible choice is simply choosing $l_\alpha \le u_\alpha$ such that
   $$F(u_\alpha) - F(l_\alpha) \ge 1 - \alpha.$$
2. Given Step 1-(b), suppose we have chosen l_α, u_α, then it holds that
   $$\mathbb{P}(l_\alpha \le g(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n; \theta) \le u_\alpha) \ge 1 - \alpha.$$

3. We can understand $l_\alpha \leq g(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n; \theta) \leq u_\alpha$ as an inequality involving variable θ, consequently, let us solve this inequality in terms of θ, which would typically gives us an interval on θ with endpoints depending on $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, l_α and u_α, that is

$$\boldsymbol{L}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n; u_\alpha, l_\alpha) \leq \theta \leq \boldsymbol{U}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n; u_\alpha, l_\alpha). \tag{1.2}$$

4. In other words, the following two events defined by random samples $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ are the same:

$$\begin{aligned} &\{(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) : l_\alpha \leq g(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n; \theta) \leq u_\alpha\} \\ =&\{(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) : \boldsymbol{L}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n; u_\alpha, l_\alpha) \leq \theta \leq \boldsymbol{U}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n; u_\alpha, l_\alpha)\}. \end{aligned}$$

5. Given Step 2, we immediately have

$$\mathbb{P}(\boldsymbol{L}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n; u_\alpha, l_\alpha) \leq \theta \leq \boldsymbol{U}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n; u_\alpha, l_\alpha)) \geq 1 - \alpha.$$

6. Consequently, $[\boldsymbol{L}, \boldsymbol{U}]$ is an $(1 - \alpha)$-confidence interval for θ.

2 CI for Mean of Normal Distributions (Known Variance)

We are now ready to see the first set of CIs, that are constructed exactly by applying the procedure described in Proposition 1.1.

2.1 Two-sided CI

Example 2.1. Consider $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are I.I.D. samples from $N(\mu, \sigma^2)$ with known σ^2 and unknown μ. Given $\alpha \in (0, 1)$, how do we construct a two-sided $(1 - \alpha)$-CI for the mean μ?

Solution:

1. Let us define the following statistics that involves both the samples and the unknown parameter μ:

$$\boldsymbol{Y}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n; \mu) = \frac{\overline{\boldsymbol{X}} - \mu}{\sigma/\sqrt{n}}.$$

We will write $\boldsymbol{Y}$ in short for $\boldsymbol{Y}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n; \mu)$. Now what is the distribution of $\boldsymbol{Y}$?

2. Now, for any $\beta \in (0, 1)$, let us define z_β to be the following constant:

$$\mathbb{P}(\boldsymbol{Z} \geq z_\beta) = \beta, \text{ where } \boldsymbol{Z} \text{ follows } N(0, 1). \tag{2.1}$$

We say z_β is the β-percentage point of standard normal distribution, which can be found with the Φ-table (or Z-table) given an β. You can also find it on Canvas. It is easy to verify that

$$\mathbb{P}(\boldsymbol{Z} \leq -z_\beta) = \beta.$$

This is due to the following property on the C.D.F of standard normal distribution:

$$\Phi(x) + \Phi(-x) = 1, \text{ for any } x.$$

Consequently, we have that

$$\mathbb{P}(-z_{\alpha/2} \leq \boldsymbol{Y} \leq z_{\alpha/2}) = 1 - \alpha,$$

which is equivalent to

$$\mathbb{P}(-z_{\alpha/2} \leq \frac{\overline{\boldsymbol{X}} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}) = 1 - \alpha \tag{2.2}$$

3. Now given $-z_{\alpha/2} \leq \frac{\overline{\boldsymbol{X}}-\mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}$, let us solve for μ, which gives that

$$\overline{\boldsymbol{X}} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{\boldsymbol{X}} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

4. That is, the following two events are the same:

$$\begin{aligned}&\left\{(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) : -z_{\alpha/2} \leq \frac{\overline{\boldsymbol{X}} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right\} \\ =&\left\{(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) : \overline{\boldsymbol{X}} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{\boldsymbol{X}} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right\}.\end{aligned}$$

5. Consequently, given (4.4), we know that

$$\mathbb{P}\left(\overline{\boldsymbol{X}} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{\boldsymbol{X}} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

6. Finally, we can conclude that

$$[\overline{\boldsymbol{X}} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \leq \overline{\boldsymbol{X}} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}] \tag{2.3}$$

is a $1 - \alpha$ confidence interval for μ.

■

Exercise 2.1. Can we construct another confidence interval of μ by slightly changing one of the steps in the above construction?

Solution:

■

Hopefully, we can see from the above that the construction of CIs, similar to the construction of point estimations, can be fairly flexible. Many CIs of different forms can be constructed for the same parameters. They would have different properties. To compare different CIs, we also have to propose a meaningful metric to characterize their "accuracies" – much similar to the mean-squared error for point estimations.

Such detailed concepts are out of the scope for our discussion in this class. But in a nutshell, we can perhaps agree on a general rule that a good CI should satisfy:

Given $\alpha \in (0, 1)$ and n I.I.D. samples, a good $(1-\alpha)$-CI should be the one that has minimum width among all the possible CIs.

2.2 Large-Sample CI

Our previous construction of two-sided CI of μ assumes that the variance is known to us. In many cases this assumption is not realistic. There are two possible remedies for addressing this issue:

1. If the number of samples n is large, we can use the sample standard deviation S^2 as the proxy of σ^2.
2. If the number of samples n is low, we can use t-test to be discussed in the next section.

The first approach is formally described as follows.

Proposition 2.1 (Unknown variance, large-sample)**.** Let $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{X}_i - \overline{\boldsymbol{X}})^2$ be the sample variance. When n is large enough ($n \geq 40$), the two-sided $(1-\alpha)$-CI for μ is given by

$$[\overline{\boldsymbol{X}} - z_{\alpha/2}\frac{S}{\sqrt{n}}, \leq \overline{\boldsymbol{X}} + z_{\alpha/2}\frac{S}{\sqrt{n}}].$$

Proof. This is not formal proof – but to introduce the essential reasoning behind why can we replace the population standard deviation by its sample counterpart.

The proof simply changes one step in the construction of CI with known variance (Example 2.1).

□

2.3 Number of Samples

Previously we have seen that the $(1-\alpha)$-confidence interval for the mean of normal distribution with known variance is given by

$$[\overline{\boldsymbol{X}} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \leq \overline{\boldsymbol{X}} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}].$$

By the definition of confidence interval, this would mean that when using the sample mean $\overline{\boldsymbol{X}}$ as the point estimation of μ, the error $|\overline{\boldsymbol{X}} - \mu|$ satisfies

$$|\overline{\boldsymbol{X}} - \mu| \leq z_{\alpha/2}\frac{\sigma}{\sqrt{n}},$$

with probability at least $1-\alpha$. Now consider the following proposition.

Proposition 2.2. For any $\alpha \in (0,1)$ and any $\epsilon > 0$, to make the error $|\overline{\boldsymbol{X}} - \mu|$ smaller than ϵ with probability at least $1-\alpha$, the number of I.I.D. samples should satisfy

$$n \geq \lceil \Big(\frac{z_{\alpha/2}\sigma}{\epsilon}\Big)^2 \rceil.$$

Proof.

□

Before we continue, let us consider the following example involving the interpretation of the confidence interval.

Exercise 2.2. Consider the two-sided confidence interval in (4.5). What is the length of the interval? How does the length changes when we vary the sample size n, and when we vary the confidence level $(1-\alpha)$? Is it intuitive?

Solution:

■

2.4 One-sided CI

Previously we have discussed how to construct two-sided CIs. In many situations, we may only require a one-sided CI that takes the form of $(-\infty, \boldsymbol{U}]$ or $[\boldsymbol{L}, \infty)$, and the question is to how to construct the endpoint $\boldsymbol{U}$ or $\boldsymbol{L}$ from the I.I.D. samples. We now consider this problem in the context of the mean of normal distribution.

Example 2.2. Consider $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are I.I.D. samples from $N(\mu, \sigma^2)$ with known σ^2 and unknown μ. Given $\alpha \in (0,1)$, how do we construct a one-sided $(1-\alpha)$-CI for the mean μ. In particular, what is the upper-bounded one-sided CI $(-\infty, \boldsymbol{U}]$ with confidence $(1-\alpha)$? What is the lower-bounded one-sided CI $[\boldsymbol{L}, \infty)$ with confidence $(1-\alpha)$?

Solution: Hint: Can we generalize the approach we take in Example 2.1 to the problem considered here?

■

The following exercise will confirm a basic intuition of the upper/lower-bounded one-sided CI.

Exercise 2.3. Consider $0 < \alpha_1 < \alpha_2 < 1$. Let $(-\infty, \boldsymbol{U}_{\alpha_1}]$ be the upper-bounded one-sided CI with confidence $1 - \alpha_1$, and $(-\infty, \boldsymbol{U}_{\alpha_2}]$ be similarly defined.

Question: what is the relationship between $\boldsymbol{U}_{\alpha_1}$ and $\boldsymbol{U}_{\alpha_2}$? How does $\boldsymbol{U}_{\alpha_1}$ and $\boldsymbol{U}_{\alpha_2}$ behaves when we vary the number of I.I.D. samples n?

Solution:

■

3 CI for Mean of Normal Distributions (Unknown Variance)

Previously in Section 2.2, we have discussed how to cope with unknown variance when constructing the CI for the mean. A critical assumption assumed therein is that we have enough number of samples so that the sample variance is a good proxy of the population variance. Now what if we simply do not have enough samples? (e.g., $n \leq 20$).

3.1 Small-Sample CI with t-distribution

Definition 3.1. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be I.I.D. samples from a normal distribution with mean μ and variance σ^2. Let S^2 denote the sample variance: $S^2 = \frac{1}{n-1} \sum_{i=1}^n \left(\boldsymbol{X}_i - \overline{\boldsymbol{X}}\right)^2$. Define random variable:

$$\boldsymbol{T}_{n-1} = \frac{\overline{\boldsymbol{X}} - \mu}{S^2/\sqrt{n}},$$

then $\boldsymbol{T}_{n-1}$ is said to follow t-distribution with $n-1$ degrees of freedom. Notably, the t-distribution with $n-1$ degrees of freedom does not depend on (μ, σ^2), which are both unknown. Consequently, the distribution of $\boldsymbol{T}_{n-1}$ does not depend on the unknown (μ, σ^2), which makes it a natural statistic for constructing CI (it serves the role of the statistic $\boldsymbol{Y}$ in Proposition 1.1).

The concrete form of the density function of t-distribution with k degrees of freedom is given by

$$f(x) = \frac{\Gamma((k+1)/2)}{\sqrt{\pi k}\Gamma(k/2)} \left(\frac{x^2}{k} + 1\right)^{-\frac{k+1}{2}}.$$

It should be noted that indeed this P.D.F. does not depend on (μ, σ^2), which are both unknown. We seldom use this concrete form of P.D.F, much similar to we seldom use the P.D.F. of standard normal distribution. But the following qualitative traits of t-distribution will be very useful in understanding the connection between t-distribution and the standard normal distribution.

1. The P.D.F of t-distribution (regardless of the degrees of freedom) is symmetric, unimodal, and attains maximum at $x = 0$.

2. It has a heavier tail than the P.D.F. of standard normal distribution. The tail becomes thinner when the degrees of freedom increases. In particular, if $k \to \infty$, then the P.D.F. of t-distribution converges to that of the standard normal distribution.

For the purpose of constructing CIs, let us define the percentage point of the t-distribution, similar to the percentage point we define for the standard normal distribution (2.1).

Definition 3.2 (Percentage point of t-distribution)**.** For any $\beta \in (0, 1)$, the β-percentage point of the t-distribution with k degrees of freedom is denoted by $t_{\beta,k}$, and is defined as

$$\mathbb{P}(\boldsymbol{T}_k \geq t_{\beta,k}) = \beta. \tag{3.1}$$

With the introduction of t-distribution, let us consider constructing the two-sided CI of the mean with unknown variance. As we will see, the approach is exactly the same as we construct the two-sided CI with known variance (Example 2.1).

Example 3.1. Consider $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are I.I.D. samples from $N(\mu, \sigma^2)$ with unknown σ^2 and unknown μ. Given $\alpha \in (0, 1)$, how do we construct a two-sided $(1 - \alpha)$-CI for the mean μ?

Solution:

1. Let us define the following statistics that involves both the samples and the unknown parameter μ:

$$\boldsymbol{Y}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n; \mu) = \frac{\overline{\boldsymbol{X}} - \mu}{S/\sqrt{n}}.$$

We will write $\boldsymbol{Y}$ in short for $\boldsymbol{Y}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n; \mu)$. Now what is the distribution of $\boldsymbol{Y}$?

2. Given the Definition 3.2, we have that

$$\mathbb{P}(-t_{\alpha/2,n-1} \leq \boldsymbol{Y} \leq t_{\alpha/2,n-1}) = 1 - \alpha,$$

which is equivalent to

$$\mathbb{P}(-t_{\alpha/2,n-1} \leq \frac{\overline{\boldsymbol{X}} - \mu}{S/\sqrt{n}} \leq t_{\alpha/2,n-1}) = 1 - \alpha \tag{3.2}$$

3. Now given $-t_{\alpha/2,n-1} \leq \frac{\overline{\boldsymbol{X}}-\mu}{S/\sqrt{n}} \leq t_{\alpha/2,n-1}$, let us solve for μ, which gives that

$$\overline{\boldsymbol{X}} - z_{\alpha/2}\frac{S}{\sqrt{n}} \leq \mu \leq \overline{\boldsymbol{X}} + z_{\alpha/2}\frac{S}{\sqrt{n}}.$$

4. That is, the following two events are the same:

$$\begin{aligned} &\left\{(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) : -t_{\alpha/2,n-1} \leq \frac{\overline{\boldsymbol{X}} - \mu}{S/\sqrt{n}} \leq t_{\alpha/2,n-1}\right\} \\ =&\left\{(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) : \overline{\boldsymbol{X}} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}} \leq \mu \leq \overline{\boldsymbol{X}} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right\}. \end{aligned}$$

5. Consequently, given (4.4), we know that

$$\mathbb{P}\left(\overline{\boldsymbol{X}} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}} \leq \mu \leq \overline{\boldsymbol{X}} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right) = 1 - \alpha.$$

6. Finally, we can conclude that

$$[\overline{\boldsymbol{X}} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}, \leq \overline{\boldsymbol{X}} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}] \tag{3.3}$$

is a $1 - \alpha$ confidence interval for μ.

■

Exercise 3.1. Construct the upper/lower bounded one-sided CI for μ with unknown variance, $(1 - \alpha)$ confidence level.

Solution:

∎

4 CI for Variance of Normal Distribution

Previously we have discussed how to construct CI of the mean μ given samples from $N(\mu, \sigma^2)$. In this section, we will discuss how to construct the CI for σ^2. Notably, the construction of CI does not require any information on the sample mean μ.

Definition 4.1 (χ^2-distribution)**.** Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be I.I.D. samples from $N(\mu, \sigma^2)$. Let S^2 be the sample variance. Define random variable

$$\mathcal{X}^2 = \frac{(n-1)S^2}{\sigma^2}, \tag{4.1}$$

then we say that $\mathcal{X}^2$ follows a χ^2-distribution with $n-1$ degrees of freedom.

The probability density function (P.D.F.) of χ^2-distribution with k degrees of freedom is given by

$$f(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{(k/2)-1} e^{-x/2}, \ x > 0.$$

Again, it should be noted that χ^2-distribution with $n-1$ degrees of freedom does not depend on the unknown parameters μ or σ^2. This property makes them useful for constructing CI for σ^2 given the definition of random variable $\mathcal{X}^2$ in (4.1). The concrete form of χ^2-distribution's P.D.F. is not important to us. But the following qualitative traits might be worthy of attention:

1. The possible value of χ^2 random variable is always non-negative.

2. The P.D.F. is skewed to the right.

3. As $k \to \infty$, χ^2-distribution with k degrees of freedom converges to a normal distribution in the following sense. Let χ_k^2 be a χ^2 random variable with k degrees of freedom, then:

$$\text{The distribution of } \frac{\chi_k^2 - k}{\sqrt{2k}} \text{ converges to } N(0,1) \text{ as } k \to \infty. \tag{4.2}$$

To facilitate the construction of CI, let us define the percentage points of the χ^2-distribution.

Definition 4.2 (Percentage point of χ^2-distribution)**.** For any $\beta \in (0,1)$, the β-percentage point of the χ^2-distribution with k degrees of freedom is denoted by $\chi^2_{\beta,k}$, and is defined as

$$\mathbb{P}(\mathcal{X}_k^2 \geq \chi^2_{\beta,k}) = \beta, \tag{4.3}$$

where $\mathcal{X}_k^2$ follows the χ^2-distribution with k degrees of freedom.

With the introduction of t-distribution, let us consider constructing the two-sided CI of the mean with unknown variance. As we will see, the approach is exactly the same as we construct the two-sided CI with known variance (Example 2.1).

Example 4.1. Consider $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are I.I.D. samples from $N(\mu, \sigma^2)$ with unknown σ^2 and unknown μ. Given $\alpha \in (0, 1)$, how do we construct a two-sided $(1-\alpha)$-CI for the variance σ^2?

Solution:

1. Let us define the following statistics that involves both the samples and the unknown parameter μ:

$$\boldsymbol{Y}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n; \sigma^2) = \frac{(n-1)S^2}{\sigma^2}.$$

We will write $\boldsymbol{Y}$ in short for $\boldsymbol{Y}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n; \sigma^2)$. Now what is the distribution of $\boldsymbol{Y}$?

2. Given the Definition 4.1, we have that

$$\mathbb{P}(-\chi^2_{1-\alpha/2,n-1} \leq \boldsymbol{Y} \leq \chi^2_{\alpha/2,n-1}) = 1-\alpha,$$

which is equivalent to

$$\mathbb{P}(-\chi^2_{1-\alpha/2,n-1} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{\alpha/2,n-1}) = 1-\alpha \tag{4.4}$$

3. Now given $-\chi^2_{1-\alpha/2,n-1} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{\alpha/2,n-1}$, let us solve for σ^2, which gives that

$$\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}.$$

4. That is, the following two events are the same:

$$\left\{(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) : -\chi^2_{1-\alpha/2,n-1} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{\alpha/2,n-1}\right\}$$
$$= \left\{(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) : \frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right\}.$$

5. Consequently, given (4.4), we know that

$$\mathbb{P}\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right) = 1-\alpha.$$

6. Finally, we can conclude that

$$[\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}] \tag{4.5}$$

is a $1-\alpha$ confidence interval for σ^2.

■

Having seem several examples or CIs, can we solve the following?

Exercise 4.1. Consider $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are I.I.D. samples from $N(\mu, \sigma^2)$ with unknown σ^2 and unknown μ. Given $\alpha \in (0, 1)$, how do we construct a upper/lower bounded one-sided $(1-\alpha)$-CI for the variance σ^2?

Solution:

■

On the rigorous definitions of χ^2- and t- distributions

Our previous definition of t-distribution and χ^2-distribution, although technically correct, often differs from the classical definitions. This deviation primarily stems from that our definitions in prior discussions hide away a few things in order to keep the presentation simple.

Below, we introduce the classical definitions and discuss their relationship with the definitions used in our discussions.

Definition 4.3 (χ^2-distribution)**.** If $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_k$ are I.I.D. samples from $N(0, 1)$, then

$$Q = \sum_{i=1}^{k} \boldsymbol{Z}_i^2$$

has a χ^2-random variable with k-degrees of freedom.

Now let us consider the above definition and Definition 4.1.

Exercise 4.2. Clearly, Definition 4.1 and Definition 4.3 of χ^2-distributions are different – how could we reconcile their differences?

Before we answer Exercise 4.2, let us first consider the formal definition of t-distribution.

Definition 4.4 (t-distribution)**.** Let $\boldsymbol{Z}$ follows $N(0,1)$. Let $\boldsymbol{V}$ be a χ^2-distribution with k degrees of freedom. Suppose $\boldsymbol{Z}$ and $\boldsymbol{V}$ are independent, then

$$\boldsymbol{T} = \frac{\boldsymbol{Z}}{\sqrt{\boldsymbol{V}/k}}$$

follows a t-distribution with k degrees of freedom.

Exercise 4.3. Clearly, Definition 3.1 and Definition 4.4 of t-distributions are different – how could we reconcile their differences?

The following proposition is very useful for samples coming from the normal distributions, it may also appear to be surprising at the first glance. We omit their proofs due to the fact that both require quite a bit of computation.

Proposition 4.1. Let $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ be I.I.D. samples from $N(\mu, \sigma^2)$. Let $\overline{\boldsymbol{X}}$ be the sample mean and S^2 be the sample variance.

1. $\overline{\boldsymbol{X}}$ and S^2 are independent random variables.
2. We have $\frac{(n-1)S^2}{\sigma^2}$ follows χ^2-distribution (Definition 4.3) with $n-1$ degrees of freedom.

Now with the above proposition, we can show Exercise 4.2 and 4.3.

Solution: [Solution of Exercise 4.2 and 4.3]

■

ISyE 3770: Hypothesis Testing

To motivate our discussion in hypothesis testing, let us see in the following example, where we have a most rudimentary version of ”hypothesis testing” without rigorously defining what ”hypothesis testing” really is.

Example 0.1. We have a coin, but do not know whether it is fair or not. Let us toss the coin for 100 times, and observe that the head occurs for 70 times. Then what can we say about the fairness of the coin?

Solution: Let us adopt the following thinking process. Suppose the coin is fair, then the number of heads $\boldsymbol{X}$ simply follows Binomial distribution $B(n, p)$ with $n = 100$ and $p = 1/2$.

Then let us determine the probability of the observed event:

$$\mathbb{P}(\boldsymbol{X} \geq 70) = \sum_{k=70}^{100} \binom{100}{k} * (0.5)^{100} \approx \mathbb{P}(\boldsymbol{Z} \geq 4) = 0.00003,$$

where the second equality (where $\boldsymbol{Z}$ is standard normal r.v.) follows from the normal approximation of $B(100, 1/2)$, and the last equality follows from checking the Φ-table (i.e., the C.D.F. of standard normal distribution.

Now what can we conclude from the above calculation?

If the coin is fair, then the probability of observing 70 or more heads is at most 0.00003 – which is absurdly low.

That is, if the coin is fair, then we have observed an extremely-small-probability event happening.

Consequently, we have great confidence that the coin is not fair! ■

1 Basic Concepts

Definition 1.1 (Statistical Hypothesis)**.** A statistical hypothesis is a statement about the parameters of one or two populations.

Typically, a statistical hypothesis looks like the following. We will take the coin example for illustration.

1. Two-sided hypothesis:

$$\underbrace{H_0 : p = 1/2}_{\text{null hypothesis}}. \qquad \underbrace{H_1 : p \neq 1/2}_{\text{two-sided alternative hypothesis}} \qquad .$$

2. One-sided hypothesis (greater than):

$$\underbrace{H_0 : p = 1/2}_{\text{null hypothesis}}. \qquad \underbrace{H_1 : p > 1/2}_{\text{one-sided alternative hypothesis}}.$$

3. One-sided hypothesis (smaller than):

$$\underbrace{H_0 : p = 1/2}_{\text{null hypothesis}}. \qquad \underbrace{H_1 : p > 1/2}_{\text{one-sided alternative hypothesis}}.$$

Now suppose we have formed a statistical hypothesis in one of the above form, how do we test it? This is done typically by collecting I.I.D. samples, constructing a testing statistic, and then using the testing statistic to determine if we want to reject the null hypothesis or not. We will discuss this procedure in detail in the following sections. But in summary, a testing method (or test for simplicity) for a statistical hypothesis looks like the following.

Definition 1.2 (Typical Skeleton of a Method for Testing Hypothesis)**.**

1. Collect I.I.D. samples.

2. Determine a test statistic that is a function of the samples (hence a random variable).

3. Determine a rejection region (this region does not depend on data) that falls into the range of the test statistics (here the range is all the possible values that the test statistic can take). This rejection region typically depends on the so-called significance level $\alpha \in (0, 1)$ we aim to achieve, a concept which we will introduce shortly. For examples of how α affects the rejection region, we can refer to the following sections containing detailed testing methods.

4. If the test statistic falls into the rejection region, we say that we reject H_0 with significance level α; otherwise, we say that under significance level α, we do not reject H_0.

For any given method of testing a statistical hypothesis, there are several errors associated with such a method.

Definition 1.3 (Types of Errors for Hypothesis Testing)**.** For any method of testing a statistical hypothesis, we define

1. Type-I error: Rejecting the null hypothesis H_0 when it is true is defined as a type I error. In addition, we define

   $$\alpha = \mathbb{P}(\text{Type-I error occurs}).$$

   Type I error probability is called the significance level.

2. Type-II error: Failing to reject the null hypothesis H_0 when it is false is defined as a type II error. In addition, we define

   $$\beta = \mathbb{P}(\text{Type-II error occurs}).$$

In summary, we have the following:

	H_0 is true	**H_0 is false**
Fail to reject H_0	No error	Type II error
Reject H_0	Type I error	No error

Remark 1.1. It should be noted that rejection of the null hypothesis H_0 as a strong conclusion. In contrast, accepting H_0 is treated as a weak conclusion – and rather than saying we "accept H_0," we prefer the terminology "fail to reject H_0." A useful analog exists between hypothesis testing and a jury trial. In a trial, the defendant is assumed innocent (this is like assuming the null hypothesis to be true). If strong evidence is found to the contrary, the defendant is declared to be guilty (we reject the null hypothesis). If evidence is insufficient, the defendant is declared to be not guilty. This is not the same as proving the defendant innocent and so, like failing to reject the null hypothesis, it is a weak conclusion.

Definition 1.4 (Power of a test)**.** For a method of testing a statistical hypothesis, its power is given by $1 - \beta$. That is

$$\text{Power} = \mathbb{P}(\text{Reject } H_0 \text{ when } H_0 \text{ is is false}).$$

Remark 1.2. Typically, there is a tradeoff between α-error (significance level) and β-error (1- power). Consider an extreme case. Suppose we just designed a trivial testing method that would do the following: do not reject H_0 no matter what samples we have observed – then $\alpha = 0$. On the other hand, by doing so we would have $\beta = 1$ (since whenever H_0 is false we still fails to reject it) – consequently, the test would have 0-power.

In practice, we would like a test that seeks the optimal balance between the α-error (significance level) and β-error (1- power). Typically, we seek the following:

Fixing the significance level α, find the testing method (or test for simplicity), that achieves maximum power (or equivalently, minimal β-error).

Finding an "optimal" test method that meets the above criterion is out of the scope of our discussions in this class and is one of the central topic in mathematical statistics. Nevertheless, typical statistical procedures we would encounter in the following subsections typically have a nice balance of the above tradeoff - if not optimal.

Definition 1.5 (P-value)**.** Suppose the null hypothesis H_0 is true, the probability of obtaining a test statistic (random variable) at least as extreme as the one obtained computed from data (collected I.I.D. samples).

Remark 1.3. When describing the the testing method, we do not need the concrete values of collected data. However, when computing the P-value of a test method, we need concrete values of collected data. See Example 2.2 in the next section for illustration.

Remark 1.4. Depending on whether the alternative hypothesis H_1 is one-sided or two-sided, the definition of "extreme" is different, consequently the computation of P-value is also different. We will introduce the computation of P-values when we consider more concrete examples of testing methods. We will also introduce the usage of P-values when discussing detailed examples.

Many of our discussions in this section seem fairly abstract. In the following sections, we will introduce examples that will make every notion introduced in this section concrete.

2 Testing the mean of normal distributions – variance known

2.1 Two-sided Test and Application of P-value

Example 2.1 (Two-sided Test)**.** Suppose we have $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ following from $N(\mu, \sigma^2)$ with unknown μ and known σ^2. Suppose we want to test a statistical hypothesis that whether the mean of the normal distribution equals a given number μ_0:

$$H_0 : \mu = \mu_0, \;\; H_1 : \mu \neq \mu_0.$$

Step 1: Let us consider the following test statistics:

$$\boldsymbol{Z}_0 = \frac{\overline{\boldsymbol{X}} - \mu_0}{\sigma/\sqrt{n}}. \tag{2.1}$$

Now here is the immediate question, what is the distribution of $\boldsymbol{Z}_0$ if H_0 is true? This question is almost immediate given our discussion in the chapter of statistical intervals:

$$\boldsymbol{Z}_0 \text{ follows } N(0,1) \text{ if } H_0 \text{ is true.} \tag{2.2}$$

Step 2: Construct rejection region. To do so, recall that our rejection region should be constructed in a way such that the Type I error probability is at most α. Moreover, since our alternative hypothesis H_1 is two-sided, our rejection area should look like the following form:

$$\mathcal{R}_\alpha = (-\infty, -R_\alpha] \cup [R_\alpha, \infty) \text{ for some } R_\alpha > 0.$$

Such a form of rejection region essentially states the following: "if the testing statistic deviates from 0 with a large quantity (at least R_α), then we should reject H_0 at significance level α".

To make Type I error probability being at most α, we require

$$\mathbb{P}(\boldsymbol{Z}_0 \in \mathcal{R}_\alpha, \text{ while } H_0 \text{ is true}) \leq \alpha. \tag{2.3}$$

It remains to determine the choice of R_α for a given significance level $\alpha > 0$. To satisfy (2.3), we can simply choose $R_\alpha = z_{\alpha/2}$ and let

$$\mathcal{R}_\alpha = (-\infty, -z_{\alpha/2}] \cup [z_{\alpha/2}, \infty), \tag{2.4}$$

where $z_{\alpha/2}$ is the $(\alpha/2)$-percentage point of standard normal distribution (defined in Chapter 8 – Statistical Intervals). Indeed, we have

$$\mathbb{P}(\boldsymbol{Z}_0 \in \mathcal{R}_\alpha, \text{ while } H_0 \text{ is true}) = \mathbb{P}(\boldsymbol{Z} \in \mathcal{R}_\alpha) = \alpha, \tag{2.5}$$

where in the first equality $\boldsymbol{Z}$ is a standard normal random variable given (2.2).

Step 3. Now we have essentially defined a statistical test that looks like this:

1. Compute $\boldsymbol{Z}_0 = \frac{\overline{\boldsymbol{X}} - \mu_0}{\sigma/\sqrt{n}}$ from the I.I.D. samples.
2. For $\alpha \in (0,1)$, if $\boldsymbol{Z}_0 \in [-z_{\alpha/2}, z_{\alpha/2}]$, we say we do not reject H_0 at the significance level α. On the other hand, if $\boldsymbol{Z}_0 < -z_{\alpha/2}$ or $\boldsymbol{Z}_0 > z_{\alpha/2}$, we say we reject the null hypothesis H_0 at the significance level α.
3. Given (2.5), the significance level of the above test is indeed α.

Let us visualize the rejection and acceptance region of the above test with a significance level α.

Now let us consider the P-value of the previously designed testing method.

Example 2.2 (Computation and Application of P-value)**.** Suppose we know $\sigma = 1$, and we want to test if $\mu = \mu_0 = 1.0$ with two-sided test in Example 2.1. We have collected 10 I.I.D. samples, with values being $0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$.

What is the P-value of the test for Example 2.1?

Using the P-value, can we reject the null hypothesis at significance level 0.05? Can we reject the null hypothesis at significant level 0.1?

Solution: Given the 10 samples, the test statistics defined in 2.1 is

$$\boldsymbol{Z}_0 = (0.45 - 1.0) \cdot \sqrt{10} = -1.74.$$

Consider the definition of P-values in Definition 1.5. Because the alternative hypothesis in Example 2.1 is two-sided, the "test statistic ($\boldsymbol{Z}_0$) being at least as extreme as the one computed from the data (-1.74)" is defined as the event

$$\{\boldsymbol{Z}_0 \leq -1.74 \text{ or } \boldsymbol{Z}_0 \geq 1.74\}$$

Since the distribution of $\boldsymbol{Z}_0$ under the null hypothsis is given by the standard normal distribution, given Definition 1.5, we have

$$\begin{aligned} \text{P-value} &= \mathbb{P}(\boldsymbol{Z}_0 \leq -1.74 \text{ or } \boldsymbol{Z}_0 \geq 1.74, \text{while } H_0 \text{ is true}) \\ &= 2 \cdot (1 - \Phi(1.74)) = 2 \cdot (1 - 0.97) = 0.06. \end{aligned}$$

Now using the P-value, let us answer the following questions:

1. Can we reject the null hypothesis at significance level 0.05?

2. Can we reject the null hypothesis at significant level 0.1?

For the first question, we make the following immediate observation: the rejection region of the two-sided test with significance level 0.05 must be something like $(-\infty, -R_{0.05}]$ and $[R_{0.05}, \infty)$ with $R_{0.05} > 1.74$. This is because of the following:

$$\begin{aligned} &\mathbb{P}(\boldsymbol{Z}_0 \leq -1.74 \text{ or } \boldsymbol{Z}_0 \geq 1.74, \text{while } H_0 \text{ is true}) = 0.06, \\ \text{while} \quad &\mathbb{P}(\boldsymbol{Z}_0 \leq -R_{0.05} \text{ or } \boldsymbol{Z}_0 \geq R_{0.05}, \text{while } H_0 \text{ is true}) = 0.05. \end{aligned}$$

Consequently, the given the 10 collected samples the computed test statistics -1.74 falls out of the rejection region of the two-sided test with significance level 0.05. Thus, we do not reject the null hypothesis with a significance level 0.05.

For the second question, with the exactly the same reasoning, we can easily see that the computed test statistic -1.74 falls into the rejection region of the two-sided test with significance level 0.1. Thus, we reject the null hypothesis with a significance level 0.1.

As we can see, the computation of the P-value for a test helps us conveniently do the following:

1. Given collected sample, let us compute the concrete value of the test statistics, from which we can determine the P-value (we have seen how to do this for two-sided test, we will also see how to do this for one-sided test).

2. Now for a given significant level α, we do the following:

   If P-value $< \alpha$, we reject the null hypothesis with significance level α.

   This is because in this case, one can easily reason that the computed test statistics falls into the rejection region of the test with significance level α – just as what we did before.

3. On the other hand,

   If P-value $\geq \alpha$, we do not reject the null hypothesis with significance level α.

   This is because in this case, one can easily reason that the computed test statistics falls into the acceptance region of the test with significance level α – just as what we did before.

■

Hopefully, with the above examples of two-sided test and the application of P-values. We have a better understanding on the basic concepts of statistical hypothesis testing in Section 1.1.

Although we will continue to see several examples of hypothesis testing, but in a nutshell the essential idea has been introduced with the previous two examples. In principle, we will be able to derive a statistical test for other scenarios we will consider, with just the same reasoning we have seen in the previous two examples. I would suggest to really make sure we understand these examples before we continue to see other examples of hypothesis testing.

We have not touched upon the power (or equivalently the β-error) of a test, which we will postpone a bit later.

Let us proceed to the discussion of the one-sided test – we will see that all the previous ideas of two-sided test extends in an almost trivial fashion to the one-sided test.

2.2 One-sided Test and Application of P-value

As we will see, performing the one-sided test is almost the same as the two-sided test.

Example 2.3 (Two-sided Test)**.** Suppose we have $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ following from $N(\mu, \sigma^2)$ with unknown μ and known σ^2. Suppose we want to test a statistical hypothesis that whether the mean of the normal distribution equals a given number μ_0:

$$H_0 : \mu = \mu_0, \;\; H_1 : \mu > \mu_0.$$

Step 1: Let us consider the following test statistics:

$$\boldsymbol{Z}_0 = \frac{\overline{\boldsymbol{X}} - \mu_0}{\sigma/\sqrt{n}}. \tag{2.6}$$

Again, we have

$$\boldsymbol{Z}_0 \text{ follows } N(0,1) \text{ if } H_0 \text{ is true.} \tag{2.7}$$

Step 2: Construct rejection region. To do so, recall that our rejection region should be constructed in a way such that the Type I error probability is at most α. Moreover, since our alternative hypothesis H_1 is one-sided and is in the ">" form, our rejection area should look like the following form:

$$\mathcal{R}_\alpha = [R_\alpha, \infty) \text{ for some } R_\alpha > 0.$$

Such a form of rejection region essentially states the following: "if the testing statistic deviates is greater than 0 with a large quantity (at least R_α), then we should reject H_0 at significance level α".

To make Type I error probability being at most α, we require

$$\mathbb{P}(\boldsymbol{Z}_0 \in \mathcal{R}_\alpha, \text{ while } H_0 \text{ is true}) \leq \alpha. \tag{2.8}$$

It remains to determine the choice of R_α for a given significance level $\alpha > 0$. To satisfy (2.8), we can simply choose $R_\alpha = z_\alpha$ and let

$$\mathcal{R}_\alpha = [z_\alpha, \infty), \tag{2.9}$$

where z_α is the α-percentage point of standard normal distribution (defined in Chapter 8 – Statistical Intervals). Indeed, we have

$$\mathbb{P}(\boldsymbol{Z}_0 \in \mathcal{R}_\alpha, \text{ while } H_0 \text{ is true}) = \mathbb{P}(\boldsymbol{Z} \in \mathcal{R}_\alpha) = \alpha, \tag{2.10}$$

where in the first equality $\boldsymbol{Z}$ is a standard normal random variable given (2.7).

Step 3. Now we have essentially defined a statistical test that looks like this:

1. Compute $\boldsymbol{Z}_0 = \frac{\overline{\boldsymbol{X}} - \mu_0}{\sigma/\sqrt{n}}$ from the I.I.D. samples.
2. For $\alpha \in (0, 1)$, if $\boldsymbol{Z}_0 \leq z_\alpha$, we say we do not reject H_0 at the significance level α. On the other hand, if $\boldsymbol{Z}_0 > z_\alpha$, we say we reject the null hypothesis H_0 at the significance level α.
3. Given (2.10), the significance level of the above test is indeed α.

Let us visualize the rejection and acceptance region of the above test with a significance level α.

Now let us consider the P-value of the previously designed testing method.

Example 2.4 (Computation and Application of P-value)**.** Suppose we know $\sigma = 1$, and we want to test if $\mu = \mu_0 = 1.0$ with one-sided test in Example 2.3. We have collected 10 I.I.D. samples, with values being $0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$.

What is the P-value of the test for Example 2.3?

Using the P-value, can we reject the null hypothesis at significance level 0.05? Can we reject the null hypothesis at significant level 0.1?

Solution: Given the 10 samples, the test statistics defined in Example 2.3 is

$$\boldsymbol{Z}_0 = (0.45 - 1.0) \cdot \sqrt{10} = -1.74.$$

Consider the definition of P-values in Definition 1.5. Because the alternative hypothesis in Example 2.1 is one-sided and in the ">" form, the "test statistic ($\boldsymbol{Z}_0$) being at least as extreme as the one computed from the data (-1.74)" is defined as the event

$$\{\boldsymbol{Z}_0 \geq -1.74\}$$

Since the distribution of $\boldsymbol{Z}_0$ under the null hypothsis is given by the standard normal distribution, given Definition 1.5, we have

$$\text{P-value} = \mathbb{P}(\boldsymbol{Z}_0 \geq -1.74, \text{while } H_0 \text{ is true}) = 1 - \Phi(-1.74) = 0.96.$$

Now using the P-value, let us answer the following questions:

1. Can we reject the null hypothesis at significance level 0.05?
2. Can we reject the null hypothesis at significant level 0.99?

For the first question, we make the following immediate observation: the rejection region of the one-sided test with significance level 0.05 must be something like $[R_{0.05}, \infty)$ with $R_{0.05} > -1.74$. This is because of the following:

$$\begin{aligned} \mathbb{P}(\boldsymbol{Z}_0 \geq -1.74, \text{while } H_0 \text{ is true}) &= 0.96, \\ \text{while} \quad \mathbb{P}(\boldsymbol{Z}_0 \geq R_{0.05}, \text{while } H_0 \text{ is true}) &= 0.05. \end{aligned}$$

Consequently, the given the 10 collected samples the computed test statistics -1.74 falls out of the rejection region of the two-sided test with significance level 0.05. Thus, we do not reject the null hypothesis with a significance level 0.05.

For the second question, with the exactly the same reasoning, we can easily see that the computed test statistic -1.74 falls into the rejection region of the two-sided test with significance level 0.99. Thus, we reject the null hypothesis with a significance level 0.99.

As we can see, the computation of the P-value for a test helps us conveniently do the following:

1. Given collected sample, let us compute the concrete value of the test statistics, from which we can determine the P-value (we have seen how to do this for two-sided test, we now have seen how to do this for one-sided test).

2. Now for a given significant level α, we do the following:

   If P-value $< \alpha$, we reject the null hypothesis with significance level α.

   This is because in this case, one can easily reason that the computed test statistics falls into the rejection region of the test with significance level α – just as what we did before.

3. On the other hand,

   If P-value $\geq \alpha$, we do not reject the null hypothesis with significance level α.

   This is because in this case, one can easily reason that the computed test statistics falls into the acceptance region of the test with significance level α – just as what we did before.

■

Let us summarize our discussions of the previous examples into the following proposition.

Proposition 2.1. For testing the mean μ of $N(\mu, \sigma^2)$ with a known σ. The testing statistic is given by

$$\boldsymbol{Z}_0 = \frac{\overline{\boldsymbol{X}} - \mu_0}{\sigma/\sqrt{n}}.$$

I (construct rejection region): Depending on the form of alternative hypothesis, the rejection region for a given significance level $\alpha \in (0,1)$ is given by

1. Two-sided:

   $$H_0 : \mu = \mu_0, \; H_1 : \mu \neq \mu_0.$$

   We have rejection region

   $$\mathcal{R}_\alpha = (-\infty, -z_{\alpha/2}] \cup [z_{\alpha/2}, \infty).$$

2. One-sided (”>”):

   $$H_0 : \mu = \mu_0, \; H_1 : \mu > \mu_0.$$

   We have rejection region

   $$\mathcal{R}_\alpha = [z_\alpha, \infty).$$

3. One-sided (”<”):

   $$H_0 : \mu = \mu_0, \; H_1 : \mu < \mu_0.$$

   We have rejection region

   $$\mathcal{R}_\alpha = (-\infty, -z_\alpha].$$

II (compute P-value): Suppose we have collected samples $\boldsymbol{X}_1 = x_1, \ldots, \boldsymbol{X}_n = x_n$, and consequently computed the value of the testing statistic

$$\mathbf{z} = \boldsymbol{Z}_0(x_1, \ldots, x_n).$$

The P-value can be computed as the following:

1. Two-sided:

$$H_0 : \mu = \mu_0,\ H_1 : \mu \neq \mu_0. \ \Rightarrow\ \text{P-value} = 2 \cdot (1 - \Phi(|\mathbf{z}|)).$$

2. One-sided (">"):

$$H_0 : \mu = \mu_0,\ H_1 : \mu > \mu_0. \ \Rightarrow\ \text{P-value} = 1 - \Phi(\mathbf{z}).$$

3. One-sided ("<"):

$$H_0 : \mu = \mu_0,\ H_1 : \mu < \mu_0. \ \Rightarrow\ \text{P-value} = \Phi(\mathbf{z}).$$

III (application of P-value): Regardless of the form of the hypothesis, we always do the following:

1. If P-value $< \alpha$, we reject the null hypothesis with significance level α.

2. If P-value $\geq \alpha$, we do not reject the null hypothesis with significance level α.

3 Testing the mean of normal distributions – variance unknown

Before we begin any technical discussion of this section. Let us make the following important observation:

Our previous discussion has introduced every essential element of constructing a test method for statistical hypothesis:

1. Decide a testing statistic.

2. Forming the rejection region based on the targeted significance level α.

The essential requirement of the testing statistic is to guarantee that its distribution is known to us when the null hypothesis H_0 is true.

Now let us consider the same problem of testing the unknown μ of $N(\mu, \sigma^2)$, but with σ^2 being unknown. The null hypothesis is still the same – given a μ_0:

$$H_0 : \mu = \mu_0.$$

Previously, when σ is known, we know the following statistic

$$\boldsymbol{Z}_0 = \frac{\overline{\boldsymbol{X}} - \mu_0}{\sigma/\sqrt{n}}$$

follows the standard normal distribution.

Now since the σ is unknown, let us denote S^2 as the sample variance. Then given our discussion in the chapter of confidence interval, we know that

$$\boldsymbol{Z}_t = \frac{\overline{\boldsymbol{X}} - \mu_0}{S/\sqrt{n}}$$

will follow the t-distribution with $n-1$ degrees of freedom, if the null hypothesis H_0 is true.

Now since we know the distribution of testing statistic under the null distribution, we can follow essentially lines as our discussions in Section 2, and conclude with the following proposition.

Proposition 3.1. For testing the mean μ of $N(\mu, \sigma^2)$ with an unknown known σ. The testing statistic is given by

$$\boldsymbol{Z}_0 = \frac{\overline{\boldsymbol{X}} - \mu_0}{S/\sqrt{n}}.$$

I (construct rejection region): Depending on the form of alternative hypothesis, the rejection region for a given significance level $\alpha \in (0,1)$ is given by

1. Two-sided:

   $$H_0 : \mu = \mu_0, \ \ H_1 : \mu \neq \mu_0.$$

   We have rejection region

   $$\mathcal{R}_\alpha = (-\infty, -t_{\alpha/2,n-1}] \cup [t_{\alpha/2,n-1}, \infty),$$

   where $t_{\alpha/2,n-1}$ is the $(\alpha/2)$-percentage point of t-distribution with $n-1$ degrees of freedom, which we define in the chapter of statistical interval.

2. One-sided (">"):

   $$H_0 : \mu = \mu_0, \ \ H_1 : \mu > \mu_0.$$

   We have rejection region

   $$\mathcal{R}_\alpha = [t_{\alpha,n-1}, \infty).$$

3. One-sided ("<"):

   $$H_0 : \mu = \mu_0, \ \ H_1 : \mu < \mu_0.$$

   We have rejection region

   $$\mathcal{R}_\alpha = (-\infty, -t_{\alpha,n-1}].$$

II (compute P-value): Suppose we have collected samples $\boldsymbol{X}_1 = x_1, \ldots, \boldsymbol{X}_n = x_n$, and consequently computed the value of the testing statistic

$$\mathbf{z} = \boldsymbol{Z}_t(x_1, \ldots, x_n).$$

The P-value can be computed as the following:

1. Two-sided:

   $$H_0 : \mu = \mu_0,\ H_1 : \mu \neq \mu_0. \ \Rightarrow\ \text{P-value} = 2 \cdot (1 - \Psi_{n-1}(|\mathbf{z}|)),$$

   where $\Psi_{n-1}(z)$ is the cumulative distribution of the t-distribution with $n-1$ degrees of freedom (much similar to the $\Phi(z)$ function, that is the cumulative distribution function of standard normal distribution).

2. One-sided (">"):

   $$H_0 : \mu = \mu_0,\ H_1 : \mu > \mu_0. \ \Rightarrow\ \text{P-value} = 1 - \Psi_{n-1}(\mathbf{z}).$$

3. One-sided ("<"):

   $$H_0 : \mu = \mu_0,\ H_1 : \mu < \mu_0. \ \Rightarrow\ \text{P-value} = \Psi_{n-1}(\mathbf{z}).$$

III (application of P-value): Regardless of the form of the hypothesis, we always do the following:

1. If P-value $< \alpha$, we reject the null hypothesis with significance level α.

2. If P-value $\geq \alpha$, we do not reject the null hypothesis with significance level α.

4 General Hypothesis Testing

Our previous two sections have discussed how to perform hypothesis testing for the mean of the normal distributions. Although the detailed testing statistics and the rejection region is specialized for the concrete problem considered there – the thinking process if completely general.

In conclusion, to construct a test method for a given statistical hypothesis, we need to

1. Identify the testing statistic $\boldsymbol{Z}_{\text{test}}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ so that we know its distribution when the null hypothesis holds true.

2. Determine the rejection region based on the targeted significance level α. The form of the rejection region also depends on the form of the alternative hypothesis ("$=$ / $>$ / $<$").

Once we have collected detailed samples $\boldsymbol{X}_1 = x_1, \ldots, \boldsymbol{X}_n = x_n$, we can also compute the P-value of a given test method and apply the P-value as follows:

1. Given the form of the alternative hypothesis ("$=$ / $>$ / $<$"), determine the event that "test statistic ($\boldsymbol{Z}_0$) being at least as extreme as the one computed from the data".

2. Evaluate the probability of such an event assuming that null hypothesis is true $\Rightarrow$ this gives the P-value.

3. For any $\alpha \in (0, 1)$,

   (a) If P-value $\geq \alpha$, then we do not reject the null hypothesis with significance level α.

   (b) If P-value $< \alpha$, then we reject the null hypothesis with significance level α.