Introduction to Probability Theory

Math 356 Course Notes

McGill University

Jiajun Zhang

June 23, 2025

Acknowledgments

I would like to extend my deepest thanks and appreciation to the following people,
without whose support this note would not have been possible:

[Masoud Asgharian], *Professor, McGill University*
I would like to express my deepest gratitude to Professor Masoud,
the instructor for this course, whose guidance and expertise
were invaluable throughout the development of my notes.

Despite all efforts, there may still be some typos, unclear explanations, etc. If you find potential mistakes, or any suggestions regarding concepts or formats, etc., feel free to reach out to the author at *zhangjohnson729@gmail.com*.

Contents

Chapter 1

Basic Introduction to probability

σ-Algebra

We begin with some definitions:

Definition

Definition 1. *Let $\{A_n\}$ be a sequence of sets, the set of all points $\omega \in \Omega$ (where Ω is the reference set) that belong to A_n for infinitely many values of n is known as the limit supremum of the sequence and is defined by*

$$\lim_{n\to+\infty} \sup A_n \quad or \quad \overline{\lim_{n\to+\infty}} A_n$$

Similarly, the set of all points that belong to A_n for all but a finite number of values of n is known as the limit inferior of the sequence $\{A_n\}$ and is denoted by

$$\lim_{n\to+\infty} \inf A_n \quad or \quad \underline{\lim_{n\to+\infty}} A_n$$

If $\lim_{n\to+\infty} \sup A_n = \lim_{n\to+\infty} \inf A_n$, we say the limit exists and denote $\lim_{n\to+\infty} A_n$ to be its limit.

Corollary

Corollary 1.

$$\underline{\lim_{n\to+\infty}} A_n = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k \subseteq \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k = \overline{\lim_{n\to+\infty}} A_n$$

Given the definition of lim sup and lim inf last lecture, we now give some examples:

① Let $A_n = \{n\}, n \in \mathbb{N}$, then by definition we have

$$\underline{\lim_{n\to\infty}} A_n = \varnothing \quad \text{and} \quad \overline{\lim_{n\to\infty}} A_n = \varnothing$$

② Let $A_n = \{(-1)^n\}, n \in \mathbb{N}$, then by definition we have

$$\underline{\lim_{n\to\infty}} A_n = \varnothing \quad \text{and} \quad \overline{\lim_{n\to\infty}} A_n = \{-1, 1\}$$

Definition

Definition 2. ① *If $A_n \subseteq A_{n+1}, n \in \mathbb{N}$, we say $\{A_n\}$ is non-decreasing, then $\lim_{n\to\infty} A_n = \bigcup_{n=1}^{\infty} A_n$.*

② *If $A_{n+1} \subseteq A_n, n \in \mathbb{N}$, we say $\{A_n\}$ is non-increasing, then $\lim_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} A_n$.*

Definition

Definition 3. *Let X be a space (i.e a non-empty set) and $\mathscr{F}$ be a collection of subsets of X (here $\mathscr{F}$ is the collection of subsets of X which we are going to measure). $\mathscr{F}$ is called a σ-algebra of subsets of X if:*

① $X \in \mathscr{F}$

② *If $A \in \mathscr{F}$, then $A^C := X \backslash A \in \mathscr{F}$. (Closed Under Taking Complement)*

③ *If a series of subsets $\{A_n, n \geq 1\} \in \mathscr{F}$, then $\bigcup_{n=1}^{\infty} A_n \in \mathscr{F}$. (Closed Under Countable Union)*

Based on the definition, the following propositions also hold.

Corollary

Corollary 2. *The definition of σ-algebra leads to:*

① $\varnothing \in \mathscr{F}$

② $X \in \mathscr{F}$

③ *If a series of subsets $\{A_n, n \geq 1\} \in \mathscr{F}$, then $\bigcap_{n=1}^{\infty} A_n \in \mathscr{F}$. (Closed Under Countable Intersection)*

④ *If $A_1, A_2, \cdots, A_N \in \mathscr{F}$, then $\bigcap_{n=1}^{N} A_n \in \mathscr{F}$ and $\bigcup_{n=1}^{N} A_n \in \mathscr{F}$. (Closed Under Taking Finite Union or Intersection)*

⑤ *If $A, B \in \mathscr{F}$, then $A \backslash B, B \backslash A \in \mathscr{F}$*

We say $\mathscr{F}_1$ is bigger than $\mathscr{F}_2$, if $\mathscr{F}_1 \subseteq \mathscr{F}_2$. In this case, for any space X, we may conclude that $\{\varnothing, X\}$ is the smallest σ-algebra and 2^X is the biggest σ-algebra. A σ-algebra can also be generated by a collection of sets.

Definition

Definition 4. *Let X be a space and $\mathscr{C}$ be a collection of subsets of X, then the σ-algebra generated by $\mathscr{C}$, denoted by $\sigma(\mathscr{C})$, is such that*

① *$\sigma(\mathscr{C})$ is itself a σ-algebra with $\mathscr{C} \subseteq \sigma(\mathscr{C})$*

② *If $\mathscr{F}'$ is a σ-algebra with $\mathscr{C} \subseteq \mathscr{F}'$, then $\sigma(\mathscr{C}) \subseteq \mathscr{F}'$. i.e $\sigma(\mathscr{C})$ is the smallest σ-algebra that is a super set of $\mathscr{C}$.*

Again, we have the following propositions:

Corollary

Corollary 3. ① $\sigma(\mathscr{C}) = \bigcap\{\mathscr{F} : \mathscr{C} \subseteq \mathscr{F}\}$ *for all σ-algebra $\mathscr{F}$*

② *If $\mathscr{C}$ itself is a σ-algebra, then $\sigma(\mathscr{C}) = \mathscr{C}$*

③ *If $\mathscr{C}_1, \mathscr{C}_2$ are 2 collections of subsets of X with $\mathscr{C}_1 \subseteq \mathscr{C}_2$, then $\sigma(\mathscr{C}_1) \subseteq \sigma(\mathscr{C}_2)$*

Borel σ-Algebra

An important example of σ-algebra on $\mathbb{R}$ (of subsets of $\mathbb{R}$) is called a *Borel σ-Algebra*.

Definition

Definition 5. *The Borel σ-algebra of $\mathbb{R}$ is defined by*

$$\mathfrak{B}_{\mathbb{R}} := \sigma(\{\textit{Open Sets of } \mathbb{R}\})$$

Also we need to know that the generator of $\mathfrak{B}_{\mathbb{R}}$ is not unique. By standard definition, we give

$$\mathfrak{B}_{\mathbb{R}} := \sigma(\{(a,b) : a,b \in \mathbb{R}, a < b\})$$

But we still have the following proposition:

Corollary

Corollary 4.

$$\begin{aligned}
&①\mathfrak{B}_{\mathbb{R}} := \sigma(\{(a,b] : a,b \in \mathbb{R}, a < b\}) \\
&②\mathfrak{B}_{\mathbb{R}} := \sigma(\{[a,b) : a,b \in \mathbb{R}, a < b\}) \\
&③\mathfrak{B}_{\mathbb{R}} := \sigma(\{[a,b] : a,b \in \mathbb{R}, a < b\}) \\
&④\mathfrak{B}_{\mathbb{R}} := \sigma(\{(-\infty,c) : c \in \mathbb{Q}) \\
&⑤\mathfrak{B}_{\mathbb{R}} := \sigma(\{(c,+\infty) : c \in \mathbb{Q})
\end{aligned}$$

I shall illustrate why those are indeed equal. Let's just take a look at ②, we need to show

$$\sigma(\{(a,b) : a,b \in \mathbb{R}, a < b\}) = \sigma(\{[a,b) : a,b \in \mathbb{R}, a < b\})$$

We could indeed prove $L.H.S \subseteq R.H.S$ and $R.H.S \subseteq L.H.S$.

• To show that $L.H.S \subseteq R.H.S$, we need to show

$$(a,b) \in \sigma(\{[a,b) : a,b \in \mathbb{R}, a < b\})$$

and we have

$$(a,b) = \bigcup_{n=1}^{\infty} \left[a + \frac{1}{n}, b\right) \in R.H.S$$

• Similarly,

$$[a,b) = \bigcap_{n=1}^{\infty} \left(a - \frac{1}{n}, b\right) \in L.H.S$$

And that's why they are indeed equal.

Definition

Definition 6. *A set G is called a Borel Set if $G \in \mathfrak{B}_{\mathbb{R}}$.*

In fact, any set produced by countable operations are also Borel Sets.

Corollary

Corollary 5. *A set with a single element (or singletons) $\{x\}$ is also a Borel Set.*

Proof. For a singleton $x \in \mathbb{R}$, we have

$$\{x\} = \bigcap_{n=1}^{\infty} \left(x - \frac{1}{n}, x + \frac{1}{n}\right).$$

■

Measures

Given a space X and a σ-algebra $\mathscr{F}$ of subsets of X, $(X,\mathscr{F})$ is called a measurable space.

Definition

Definition 7. *Given a measure space $(X,\mathscr{F})$, define*

$$\mu : \mathscr{F} \longrightarrow [0,+\infty]$$

is a non-negative set function, then μ is a measure if

① $\mu\{\varnothing\} = 0$;

② *If $\{A_n, n \geq 1\} \subseteq \mathscr{F}$ such that A_n's are (pairwise) disjoint, then*

$$\mu\left\{\bigcup_{n=1}^{\infty} A_n\right\} = \sum_{n=1}^{\infty} \mu(A_n)$$

this is known as countable additivity.

We say that μ is *finite* if $\mu(X) < +\infty$, we say that μ is a *probability measure* if $\mu(X) = 1$.

We say μ is σ-*finite* if $\exists\{A_n, n \geq 1\} \subseteq \mathscr{F}$ such that $X = \bigcup_{n=1}^{\infty} A_n$ with $\mu(A_n) < +\infty$.

We call the triple $(X,\mathscr{F},\mu)$ a *measure space*.

Here are a few examples of measures on $(\mathbb{R},\mathfrak{B}_{\mathbb{R}})$:

① A measure μ_1 given by

$$\mu_1 : \mathfrak{B}_{\mathbb{R}} \longrightarrow [0,+\infty], \text{such that } \forall A \in \mathfrak{B}_{\mathbb{R}}, \mu_1(A) = \begin{cases} |A|, \text{if } A \text{ is finite} \\ \\ +\infty, \text{otherwise} \end{cases}$$

We say μ_1 is the *counting measure*.

② Let $x_0 \in \mathbb{R}$, a measure μ_2 is such that

$$\forall A \in \mathfrak{B}_{\mathbb{R}}, \mu_2(A) = \begin{cases} 1, (x_0 \in A) \\ \\ 0, (x_0 \notin A) \end{cases}$$

We say μ_2 is the *probability measure*.

We now give some propositions for the measure space. Below, let $(X, \mathscr{F}, \mu)$ be a measure space.

Corollary

Corollary 6. *(Finite Additivity)*
If $A_1, A_2, \cdots, A_N \in \mathscr{F}$ are disjoint, then

$$\mu\left(A_1 \bigcup A_2 \bigcup \cdots \bigcup A_N\right) = \sum_{n=1}^{N} \mu(A_n).$$

Corollary

Corollary 7. *(Monotoncity)*
Given $A, B \in \mathscr{F}$, if $A \subseteq B$, then $\mu(A) \leq \mu(B)$.

Proof. Note that since $\mu(A), \mu(B)$ could both be infinite, so taking $\mu(B) - \mu(A)$ is incorrect. Indeed, we may construct $B = A \bigcup (B \backslash A)$, where

$$\mu(B) = \mu(A) + \mu(B \backslash A) \geq \mu(A).$$

■

Corollary

Corollary 8. *(Countable / Finite Subadditivity)*
If $\{A_n, n \geq 1\} \subseteq \mathscr{F}$, then

$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} \mu(A_n) \text{ and } \mu\left(\bigcup_{n=1}^{N} A_n\right) \leq \sum_{n=1}^{N} \mu(A_n)$$

Note that if A_n's are disjoint, then it should be "=".

Proof. Let $B_1 = A_1$, and $B_n := A_n \backslash \left(\bigcup_{i=1}^{n-1} A_i\right)$ for every $n \geq 2$, then we know that $\{B_n, n \geq 1\} \subseteq \mathscr{F}$ and B_n's are disjoint. More importantly, by our construction we have $\bigcup_{n=1}^{\infty} B_n = \bigcup_{n=1}^{\infty} A_n$. Thus by countable additivity, we have

$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \mu\left(\bigcup_{n=1}^{\infty} B_n\right) = \sum_{n=1}^{\infty} \mu(B_n) \leq \sum_{n=1}^{\infty} \mu(A_n)$$

■

Corollary

Corollary 9. *(Coutinuity From Below)*
Given $\{A_n, n \geq 1\} \subseteq \mathscr{F}$ *such that* $A_n \subseteq A_{n+1}$ *for every n, then*

$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} \mu(A_n)$$

Proof. Set $B_1 = A_1$, $B_n = A_n \backslash A_{n-1}$ for all $n \geq 2$. By definition we have $\{B_n, n \geq 1\} \subseteq \mathscr{F}$ is disjoint, and we have $\bigcup_{n=1}^{\infty} B_n = \bigcup_{n=1}^{\infty} A_n$; also for each $N \geq 1$, $\bigcup_{n=1}^{N} B_n = A_n$, which means

$$\begin{aligned}
\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \mu\left(\bigcup_{n=1}^{\infty} B_n\right) = \sum_{n=1}^{\infty} \mu(B_n) &= \lim_{N\to\infty} \sum_{n=1}^{N} \mu(B_n) \\
&= \lim_{N\to\infty} \mu\left(\bigcup_{n=1}^{N} B_n\right) \\
&= \lim_{n\to\infty} (A_n)
\end{aligned}$$

■

Corollary

Corollary 10. *(Continuity From Above)*
Given $\{A_n, n \geq 1\} \in \mathscr{F}$ *such that* $A_{n+1} \subseteq A_n$ *for every* $n \geq 1$, *if* $\mu(A_1) < +\infty$, *then*

$$\mu\left(\bigcap_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} \mu(A_n)$$

Proof. Set $B_n = A_1 \backslash A_n$, then we know that the sequence $\{B_n, n \geq 1\}$ is increasing, then we have

$$\bigcup_{n=1}^{\infty} B_n = A_1 \backslash \bigcap_{n=1}^{\infty} A_n$$

Thus,

$$\mu\left(A_1 \backslash \bigcap_{n=1}^{\infty} A_n\right) = \mu\left(\bigcup_{n=1}^{\infty} B_n\right) = \lim_{n\to\infty} \mu(B_n) = \lim_{n\to\infty} \mu(A \backslash A_n) \qquad (*)$$

Given that the measure of A_n is always finite, so $\mu(A_1 \backslash A_n) = \mu(A_1) - \mu(A_n)$ and

$$\mu\left(A_1 \backslash \bigcap_{n=1}^{\infty} A_n\right) = \mu(A_1) - \mu\left(\bigcap_{n=1}^{\infty} A_n\right)$$

Thus, by equation $(*)$,

$$\mu(A_1) - \mu\left(\bigcap_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} (\mu(A_1) - \mu(A_n))$$

that is

$$\mu\left(\bigcap_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} \mu(A_n)$$

■

The reason why we would introduce the definition of a σ-algebra, measure is that, probability, is also a measure.

An introduction to sample spaces and probability axioms

Definition

Definition 8. *A random (or a statistical) experiment is an experiment with the following:*

① *All outcomes of the experiment are known in advance;*

② *Any performance of the experiment results in an outcome is not known i advance;*

③ *The experiment can be repeated under identical conditions.*

Definition

Definition 9. *The sample space of a statistical experiment is a pair* $(\Omega, \mathscr{F})$ *where* Ω *is the set of all possible outcomes of the experiment and* $\mathscr{F}$ *is a* σ *field of the subsets of* Ω.

Here is an example to illustrate this: Suppose we are tossing a coin, and we denote H (head) and T (tail) to be the only two possible outcomes. Then we have $\Omega := \{H, T\}$ and $\mathscr{F} := \{\{H\}, \{T\}, \{H, T\}, \varnothing\}$. Element of Ω are called the *sample points* and any set $A \in \mathscr{F}$ is called an *event*.

Definition

Definition 10. *(Kolmogorov's Axioms)*
Let $(\Omega, \mathscr{F})$ *be a sample space and a set function* $\mathbb{P} : \mathscr{F} \longrightarrow \mathbb{R}$ *is called a probability measure (or simply probability) if it satisfies:*

① $\mathbb{P}(A) \geq 0$ *for all* $A \in \mathscr{F}$*;*

② $\mathbb{P}(\Omega) = 1$*;*

③ *Let* $\{A_j\}_1^\infty$ *be a sequence of disjoint sets where* $A_j \in \mathscr{F}$*, then* $\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

Corollary

Corollary 11. $\mathbb{P}$ *is monotone and substrative, i.e if* $A, B \in \mathscr{F}$ *with* $A \subseteq B$*, then* $\mathbb{P}(A) \leq \mathbb{P}(B)$.

Corollary

Corollary 12. *If* $A \in \mathscr{F}$*, then* $\mathbb{P}(A) = 1 - \mathbb{P}(A^C)$.

Now we will introduce an important theorem in probability:

Theorem

Theorem 1. *(The principle of inclusion and exclusion)*
Let $A_1, A_2, \cdots, A_n \in \mathscr{F}$*, then*

$$\mathbb{P}\left(\bigcup_{k=1}^{n} A_k\right) = \sum_{k=1}^{n} \mathbb{P}(A_k) - \sum_{k_1<k_2} \mathbb{P}(A_{k_1} \cap A_{k_2}) + \cdots + (-1)^{n+1} \sum_{k_1<\cdots<k_n} \mathbb{P}(A_{k_1} \cap \cdots \cap A_{k_n})$$

Proof. We will first consider some special cases, let's say $n = 2$, then we have

$$\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2) - \mathbb{P}(A_1 \cap A_2).$$

Similarly, consider $n = 3$, then we have

$$\begin{aligned}\mathbb{P}(A_1 \cup A_2 \cup A_3) = &\mathbb{P}(A_1) + \mathbb{P}(A_2) + \mathbb{P}(A_3) \\ &- \mathbb{P}(A_1 \cap A_2) - \mathbb{P}(A_2 \cap A_3) - \mathbb{P}(A_1 \cap A_3) \\ &+ \mathbb{P}(A_1 \cap A_2 \cap A_3)\end{aligned}$$

So when $n = 2$ or $n = 3$, it is already proven according to the construction we did using Venn diagram. Now we perform the induction on n, assume it holds for some number $N (N > 3)$, then say

$$\mathbb{P}\left(\bigcup_{i=1}^{N+1} A_i\right) = \mathbb{P}\left(\left[\bigcup_{i=1}^{N} A_i\right] \bigcup A_{N+1}\right)$$

Denote $\left[\bigcup_{i=1}^{N} A_i\right] = B$, then we have

$$\mathbb{P}\left(\bigcup_{i=1}^{N+1} A_i\right) = \mathbb{P}(B) + \mathbb{P}(A_{N+1}) - \mathbb{P}(B \cap A_{N+1})$$

■

The principle of inclusion and exculsion also leads us to another theorem:

Theorem

Theorem 2. *(Bonferroni's Inequality)*
Given n events $A_1, A_2, \cdots, A_n \in \mathscr{F}, (n > 1)$*, then*

$$\sum_{k=1}^{n} \mathbb{P}(A_k) - \sum_{i<j} \mathbb{P}(A_i \cap A_j) \leq \mathbb{P}\left(\sum_{k=1}^{n} A_k\right) \leq \sum_{k=1}^{n} \mathbb{P}(A_k)$$

Proof. The proof is basically the same idea as above, reader should try to prove this by their own. ■

Theorem

Theorem 3. *(Boole's Inequality)*
Suppose $A, B \in \mathscr{F}$, then

$$\mathbb{P}(A \cap B) \geq 1 - \mathbb{P}(A^C) - \mathbb{P}(B^C)$$

Proof.

$$\begin{aligned}
\mathbb{P}(A \cap B) &= 1 - \mathbb{P}((A \cap B)^C) \\
&= 1 - \mathbb{P}(A^C \cup B^C) \\
&= 1 - (\mathbb{P}(A^C) + \mathbb{P}(B^C) - \mathbb{P}(A^C \cap B^C)) \\
&\geq 1 - \mathbb{P}(A^C) - \mathbb{P}(B^C)
\end{aligned}$$

■

Corollary

Corollary 13. *Let $\{A_j\}_{j=1}^{\infty}$ be a sequence of events, then*

$$\mathbb{P}\left(\bigcap_{j=1}^{\infty} A_j\right) \geq 1 - \sum_{j=1}^{\infty} \mathbb{P}(A_j^C)$$

Proof. This proof can be achieved by induction, using the results in theorem 4. ■

Theorem

Theorem 4. *(The implicative rule)*
If $A_1, A_2, A_3 \in \mathscr{F}$ and $A_1 \cap A_2 \subseteq A_3$, i.e A_1, A_2 implies A_3, then $\mathbb{P}(A_3^C) \leq \mathbb{P}(A_1^C) + \mathbb{P}(A_2^C)$.

Proof. By monotoncity of probability, we have $\mathbb{P}(A_1 \cap A_2) \leq \mathbb{P}(A_3)$, i.e $\mathbb{P}((A_1 \cap A_2)^C) \geq \mathbb{P}(A_3^C)$, that is $\mathbb{P}(A_1^C \cap A_2^C) \geq \mathbb{P}(A_3^C)$, by inclusion and exculsion formula, we have $\mathbb{P}(A_1^C) + \mathbb{P}(A_2^C) - P(A_1^C \cap A_2^C) \geq \mathbb{P}(A_3^C)$, given that the probability is always non-negative, we then conclude that $\mathbb{P}(A_3^C) \leq \mathbb{P}(A_1^C) + \mathbb{P}(A_2^C)$. ■

Theorem

Theorem 5. *(Continuity of $\mathbb{P}$)*
Let $\{A_n\}_{n=1}^{\infty}$ be a sequence of non-decreasing events in $\mathscr{F}$, i.e $A_n \subseteq A_{n+1}$, then

$$\mathbb{P}(\lim_{n\to\infty} A_n) = \lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right)$$

Proof. Let $A = \bigcup_{j=1}^{\infty} A_j$, then A can be written as $A = A_n \bigcup \left(\bigcup_{j=n}^{\infty} A_{j+1} \backslash A_j\right)$, by σ additivity, we also

have

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}(A_n) + \sum_{j=n}^{\infty} \mathbb{P}(A_{j+1} \backslash A_j) \\ &= \mathbb{P}(A_n) + \sum_{j=n}^{\infty} \big(\mathbb{P}(A_{j+1}) - \mathbb{P}(A_j)\big)\end{aligned}$$

Now taking $n \to \infty$:

$$\begin{aligned}\mathbb{P}(A) &= \lim_{n\to\infty} \left[\mathbb{P}(A_n) + \sum_{j=n}^{\infty} (\mathbb{P}(A_{j+1}) - P(A_j))\right] \\ &= \lim_{n\to\infty} A_n + \lim_{n\to\infty} \left[\sum_{j=n}^{\infty} \mathbb{P}(A_{j+1}) - \mathbb{P}(A_j)\right] \\ &= \lim_{n\to\infty} \mathbb{P}(A_n)\end{aligned}$$

■

Theorem

Theorem 6. *(Continuity of $\mathbb{P}$)*
Let $\{A_n\}_{n=1}^{\infty}$ be a sequence of non-increasing events in $\mathscr{F}$, i.e $A_{n+1} \subseteq A_n$, then

$$\mathbb{P}(\lim_{n\to\infty} A_n) = \lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}\left(\bigcap_{n=1}^{\infty} A_n\right)$$

Proof. This can be done by taking $B_n = A_n^C$. ■

Conditional Probability

Definition

Definition 11. *Let $(\Omega, \mathscr{F})$ be the sample space, $A, H \in \Omega$, then we define $\mathbb{P}(A|H)$ to be the probability of A given that H occurs, known as the conditional probability of A given H.*

Example: We may think of rolling a fair dice: If no condition is given, then the probability that number 3 will face up is just $\frac{1}{6}$, but what if we given the condition that the number face up is odd? Then based on this assumption, the probability that number 3 will face up is $\frac{1}{3}$.

But how do we actually compute $\mathbb{P}(A|H)$? Think about this: By the construction of conditional probability, $\mathbb{P}(A|H)$ should somehow proportional to $\mathbb{P}(A \cap H)$, because $\mathbb{P}(A|H)$ is an event in $\mathbb{P}(A \cap H)$. We denote by $\mathbb{P}(A|H) = k\mathbb{P}(A \cap H)$ for some k, now consider $\mathbb{P}(H|H) = k\mathbb{P}(H \cap H)$, we conclude that $1 = k\mathbb{P}(H)$, so we have $k = \dfrac{1}{\mathbb{P}(H)}$.

Theorem

Theorem 7. *Let $(\Omega, \mathscr{F})$ be the sample space, $A, H \in \Omega$, then the conditional probability of A given H is given by*

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

If $\mathbb{P}(A|H) = \mathbb{P}(A)$, by theorem 8 we can further conclude $\mathbb{P}(A \cap H) = \mathbb{P}(A)\mathbb{P}(H)$.

Conditional probability is also a probability measure, it satisfies all the axioms of a probability measure.

Corollary

Corollary 14. *Let B, $\{H_n\}_{n=1}^{+\infty}$ be events, then*

(i) $\mathbb{P}(B|B) = 1$;

(ii) $\mathbb{P}(H|B) = 1 - \mathbb{P}(H^C|B)$;

(iii) If $H_1, H_2, \cdots$ are disjoint, then $\mathbb{P}\left(\bigcup_{n=1}^{+\infty} H_n \middle| B\right) = \sum_{i=1}^{n} \mathbb{P}(H_n|B)$.

Definition

Definition 12. *Let $(\Omega, \mathscr{F})$ be the sample space, $A, H \in \Omega$, then we say A, H are independent, if $\mathbb{P}(A \cap H) = \mathbb{P}(A)\mathbb{P}(H)$, i.e, the occurence of A does not depend on H.*

Corollary

Corollary 15. *Let $(\Omega, \mathscr{F})$ be the sample space, and $A_1, A_2, \cdots, A_n \in \Omega$, then*

$$\mathbb{P}\left(\bigcap_{j=1}^{n} A_j\right) = \prod_{i=1}^{n} \mathbb{P}\left(A_i \middle| \bigcap_{j=1}^{i-1} A_j\right)$$

Proof. We have

$$\mathbb{P}\left(\bigcap_{j=1}^{n} A_j\right) = \mathbb{P}\left(A_j \bigcap \left(\bigcap_{i=1}^{j-1} A_i\right)\right) = \mathbb{P}\left(\bigcap_{i=1}^{n-1} A_j\right) \mathbb{P}\left(A_j \middle| \bigcap_{i=1}^{j-1} A_i\right),$$

and we may perform the same step on $\mathbb{P}\left(\bigcap_{i=1}^{n-1} A_j\right)$, and we will eventually reach

$$\mathbb{P}\left(\bigcap_{j=1}^{n} A_j\right) = \prod_{i=1}^{n} \mathbb{P}\left(A_i \middle| \bigcap_{j=1}^{i-1} A_j\right)$$

■

Definition

Definition 13. *A sequence $\{H_n\}_{n=1}^{\infty}$ of events in $\mathscr{F}$ is called a partition of Ω if $H_i \cap H_j = \varnothing, i \neq j$ and $\bigcup_{n=1}^{\infty} H_n = \Omega$.*

Theorem

Theorem 8. *(Law of Total Probability)*

Let $\{H_n\}_{n=1}^{\infty}$ be a partition of Ω, $\mathbb{P}(H_i) \geq 0$, then $\forall B \in \mathscr{F}$, we have

$$\mathbb{P}(B) = \sum_{n=1}^{\infty} \mathbb{P}(B|H_n)\mathbb{P}(H_n)$$

Proof. We know that $B = B \cap \Omega = B \cap \bigcup_{n=1}^{\infty} H_n = \bigcup_{n=1}^{\infty} B \cap H_n$, thus

$$\begin{aligned}\mathbb{P}(B) &= \mathbb{P}\left(\bigcup_{n=1}^{\infty} B \cap H_n\right) \\ &= \sum_{n=1}^{\infty} \mathbb{P}(B \cap H_n), \text{since } B \cap H_n\text{'s are disjoint} \\ &= \sum_{n=1}^{\infty} \mathbb{P}(B|H_n)\mathbb{P}(H_n)\end{aligned}$$

■

Theorem

Theorem 9. *(Bayes Theorem)*
Let $\{H_n\}_{n=1}^{\infty}$ *be a partition of* Ω, $\mathbb{P}(H_n) > 0$. *Suppose* $B \in \mathscr{F}$ *with* $P(B) > 0$, *then*

$$\mathbb{P}(H_k|B) = \frac{\mathbb{P}(H_k)P(B|H_k)}{\sum_{i=1}^{\infty} P(H_i)\mathbb{P}(B|H_i)}$$

Proof. We know that

$$\begin{aligned}\mathbb{P}(H_k|B) &= \frac{\mathbb{P}(H_k \cap B)}{\mathbb{P}(B)} \\ &= \frac{\mathbb{P}(B|H_k)\mathbb{P}(H_k)}{\mathbb{P}(B)},\end{aligned}$$

and this follows from the result in theorem 9. ■

Example: A certain disease affects about 1 out of $10,000$ people. There is a test to check whether the person has the disease. The test is quite accurate, in particular we know that :

• The probability that the test result is positive (suggests that the person has the disease), given that the person does not have the disease is only 0.02;

• The probability that the test result is negative (suggests that the person does not have the disease), given that the person has the disease is only 0.01.

Now a random person gets tested for the disease and the result is positive. What is the probability that the person has the disease?

Solution : Let A to be the event that the person has the disease and we know that $\mathbb{P}(A) = 0.0001$ and $\mathbb{P}(A^C) = 0.9999$ (The person does not have the disease); Let H to be the event that the test is positive, (H^C) to be the event that the test is negative). Based on what the problem is given, we know that

$$\mathbb{P}(H|A^C) = 0.02; \mathbb{P}(H^C|A) = 0.01$$

and we aim to find $\mathbb{P}(A|H)$, thus using Bayes theorem, we have

$$\begin{aligned}\mathbb{P}(A|H) &= \frac{\mathbb{P}(H|A)\mathbb{P}(A)}{\mathbb{P}(H|A)\mathbb{P}(A)+\mathbb{P}(H|A^C)\mathbb{P}(A^C)} \\ &= \frac{(1-0.01)\times 0.0001}{(1-0.01)\times 0.0001+0.02\times(1-0.0001)} \\ &= 0.0049.\end{aligned}$$

Example: In Bob's town, it's rainy one thirds of the days. Given that it is rainy, there will be a heavy traffic with probability 0.5, and given that it is not rainy, there will be heavy traffic with probability 0.25. If it is rainy and there is heavy traffic, Bob will arrive late for work with probability 0.5. on the other hand (not rainy and no heavy traffic), then the probability of being late is reduced to 0.125. In other situations (rainy with no heavy traffic or not rainy with heavy traffic), the probability of being late is 0.25. Then pick a random day, Find:

(a) The probability that it is not raining and there is heavy traffic and Bob is not late for work;

(b) The probability that Bob is late for work;

(c) Given that Bob is late for work, the probability that it rained that day.

Solutions:

To start with, we define: R to be the event that it is rainy; T to be the event that there will be heavy traffic; L to be the event that Bob is late for work. Then we know the following tree diagram:

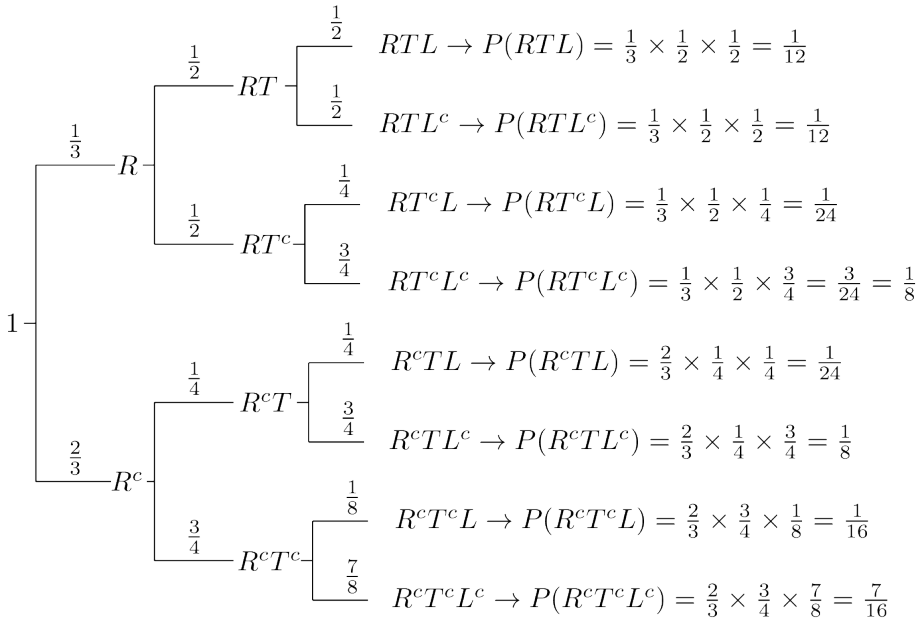


Figure 1.1: The Tree Diagram for this problem

(a) We aim to find

$$\mathbb{P}(R^C \cap T \cap L^C)$$

Then, according to the diagram we have constructed,

$$\mathbb{P}(R^C \cap T \cap L^C) = \frac{2}{3} \times \frac{1}{4} \times \frac{3}{4} = \frac{1}{8}.$$

(b) Denote $\mathbb{P}(L)$ to be the probability that Bob will be late, then we may apply the law of total probability:

$$\mathbb{P}(L) = \mathbb{P}(L|R \cap T)\mathbb{P}(R \cap T) + \mathbb{P}(L|R^c \cap T^C)\mathbb{P}(R^C \cap T^C) + \mathbb{P}(L|R^C|T)\mathbb{P}(R^C \cap T) + \mathbb{P}(L|R \cap T^C)\mathbb{P}(R \cap T^C)$$

And we are given that

$$\mathbb{P}(L|R \cap T) = \frac{1}{2}; \mathbb{P}(L|R^C \cap T^C) = \frac{1}{8}; \mathbb{P}(L|R^C \cap T) = \mathbb{P}(L|R \cap T^C) = \frac{1}{4}$$

Also we may use conditional probability to solve that

$$\begin{aligned}
\mathbb{P}(R \cap T) &= \mathbb{P}(T|R)\mathbb{P}(R) = \frac{1}{2} \times \frac{1}{3} = \frac{1}{6} \\
\mathbb{P}(R^C \cap T^C) &= \mathbb{P}(T^C|R^C)\mathbb{P}(R^C) = (1 - \mathbb{P}(T|R^C))\mathbb{P}(R^C) = \frac{3}{4} \times \frac{2}{3} = \frac{1}{2} \\
\mathbb{P}(R^C \cap T) &= \mathbb{P}(T|R^C)\mathbb{P}(R^C) = \frac{1}{4} \times \frac{2}{3} = \frac{1}{6} \\
\mathbb{P}(R \cap T^C) &= \mathbb{P}(T^C|R)\mathbb{P}(R) = (1 - \mathbb{P}(T|R))\mathbb{P}(R) = \frac{1}{2} \times \frac{1}{3} = \frac{1}{6}
\end{aligned}$$

Thus we have

$$\mathbb{P}(L) = \frac{1}{2} \times \frac{1}{6} + \frac{1}{8} \times \frac{1}{2} + \frac{1}{4} \times \frac{1}{6} + \frac{1}{4} \times \frac{1}{6} \approx 0.2292.$$

Chapter 2

Random Variables and Their Probability Distributions

Random Variables

Definition

Definition 14. *Let $(\Omega, \mathscr{F})$ be a sample space, suppose X is a finite, singled valued function that maps Ω into $\mathbb{R}$, then X is a random variable if*

$$\forall B \in \mathfrak{B}_{\mathbb{R}} : X^{-1}(B) := \{\omega \in X : X(\omega) \in B\} \in \mathscr{F}$$

That is, X is a random variable if the inverse image under X of all Borel sets in $\mathbb{R}$ are events in $\mathscr{F}$. As we mentioned earlier, we don't need to check all Borel sets in $\mathbb{R}$ (e.g open sets, closed sets, half open half closed sets), we can just check some specific sets.

Theorem

Theorem 10. *X is a random variable if and only if $\forall x \in \mathbb{R}$, we have*

$$\{\omega : X(\omega) \leq x\} \in \mathscr{F}$$

Proof. It is trivial because we know that $\mathfrak{B}_{\mathbb{R}} := \sigma((a,b])$. ■

Corollary

Corollary 16. *If X is a random variable, then $\forall a,b,c \in \mathbb{R}$, the following sets:*

$$\{X = c\}; \{a < X < b\}; \{a \leq X \leq b\}; \{a \leq X < b\}; \{X < c\}...$$

are all events in $\mathscr{F}$.

Proof. Again, it is trivial by the construction of the Borel σ field on $\mathbb{R}$. ■

Theorem

Theorem 11. *If X is a random variable, then so is $Y = aX + b$, $a,b \in \mathbb{R}$.*

Proof. For any given x, consider the set

$$\{\omega : aX(\omega) + b \leq x\}$$

i.e

$$\{\omega : aX(\omega) \leq x - b\}$$

- If $a > 0$, then we have $\{X(\omega) \leq \dfrac{x-b}{a}\} \in \mathscr{F}$;
- If $a < 0$, then we have $\{X(\omega) \geq \dfrac{x-b}{a}\} \in \mathscr{F}$;
- If $a = 0$, then then $\{\omega : aX(\omega) \leq x - b\} = \begin{cases} \{\Omega\} : x - b \geq 0 \\ \varnothing : x - b < 0 \end{cases} \in \mathscr{F}$.

Thus $Y = aX + b$ is also a random variable.

■

Here are some examples about random variables:

Example 1. For any set $A \in \mathscr{F}$, define the indicator function of set A by

$$\mathbb{I}_A(\omega) = \begin{cases} 0 : \omega \notin A \\ 1 : \omega \in A \end{cases}$$

If we consider the pre-image given by $\mathbb{I}_A$, we will find

$$\mathbb{I}_A^{-1}((-\infty, x]) = \begin{cases} \varnothing : x < 0 \\ A^C : 0 \leq x < 1 \\ \Omega : x \geq 1 \end{cases}$$

Since $\varnothing, \Omega \in \mathscr{F}$, so $\mathbb{I}_A$ is a random variable if and only if $A \in \mathbb{F}$.

Example 2. Consider tossing a fair coin two times, denote H = head face up and T = tail face up, then we know that $\Omega = \{HH, HT, TH, TT\}$ and $\mathscr{F}$ is the σ field generated by Ω, we define $X(\omega) =$ Number of H's in ω, then clearly $X(HH) = 2, X(HT) = X(TH) = 1, X(TT) = 0$, and further we have

$$X^{-1}((-\infty, x]) = \begin{cases} \varnothing : x < 0 \\ \{TT\} : 0 \leq x < 1 \\ \{TT, HT, TH\} : 1 \leq x < 2 \\ \Omega : 2 \leq x \end{cases} \in \mathscr{F}$$

and thus X is a random variable.

Corollary

Corollary 17. *If X, Y are random variables, then so is $X + Y$.*

Proof. ■

Probability Distribution of Random Variable

Definition

Definition 15. *Let $(\Omega, \mathscr{F})$ be a sample space with a probability $\mathbb{P}$, then the space $(\Omega, \mathscr{F}, \mathbb{P})$ is called a probability space.*

Theorem

Theorem 12. *The random variable X defined on the probability space $(\Omega, \mathscr{F}, \mathbb{P})$ can be mapped to another probability space $(\mathbb{R}, \mathfrak{B}_{\mathbb{R}}, Q)$ by another probability function Q defined by:*

$$\forall B \in \mathfrak{B}_{\mathbb{R}} : Q(B) = \mathbb{P}\{X^{-1}(B)\} = \mathbb{P}\{\omega : X(\omega) \in B\}$$

We call $Q = \mathbb{P}(X^{-1})$ the probability distribution of X.

Proof. To prove that Q is a probability, we just need to verify all 3 axioms from definition 7. Clearly $Q(B) \geq 0, \forall B \in \mathfrak{B}_{\mathbb{R}}$, and $Q(\mathbb{R}) = \mathbb{P}(X \in \mathbb{R}) = \mathbb{P}(\Omega) = 1$, also

$$\begin{aligned} Q\left(\sum_{i=1}^{\infty} B_i\right) = \mathbb{P}\{X^{-1}\left(\sum_{i=1}^{\infty} B_i\right)\} &= \mathbb{P}\{\sum_{i=1}^{\infty} X^{-1}(B_i)\} \\ &= \sum_{i=1}^{\infty} \mathbb{P}(X^{-1}(B_i)) = \sum_{i=1}^{\infty} Q(B_i) \end{aligned}$$

■

Definition

Definition 16. *A real valued function F defined on $\mathbb{R}$ that is non-decreasing, right continuous and satisfies*

$$F(-\infty) = 0, F(+\infty = 1)$$

is called a distribution function.

Definition

Definition 17. *Let X be a random variable defined on $(\Omega, \mathscr{F}, \mathbb{P})$, define a point function F on $\mathbb{R}$ by:*

$$\forall x \in \mathbb{R} : F(x) = Q((-\infty, x]) = \mathbb{P}\{\omega : X(\omega) \leq x\}$$

We call F the distribution function of the random variable X.

It is not hard to verify that F defined as above is indeed a distribution function.

Theorem

Theorem 13. *Given a probability Q on $(\mathbb{R}, \mathfrak{B}_{\mathbb{R}})$, there exists a distinct distribution function F such that*

$$\forall x \in \mathbb{R} : Q((-\infty, x]) = F(x)$$

Proof. Omitted. ■

Now, we will study two types of random variable and their distribution function.

Definition

Definition 18. *A random variable X defined on $(\Omega, \mathscr{F}, \mathbb{P})$ is said to be of the discrete type, if there exists a countable set $M \subseteq \mathbb{R}$ such that $\mathbb{P}(X \in M) = 1$, and the points of M which have positive values are called jump points.*

We know that all the singletons are Borel sets, thus $\{X \in E\}$ is an event.

Definition

Definition 19. *The collection of numbers $\{p_i\}$ satisfying $\mathbb{P}\{X = x_i\} = p_i \geq 0$ for all i and $\sum_i p_i = 1$ is called the probability mass function (PMF) of the random variable X, and the distribution function of X is given by*

$$F(x) = \mathbb{P}\{X \leq x\} = \sum_{x_i \leq x} p_i$$

We can use the indicator function $\mathbb{I}$ to further rewrite the definition, we now have

$$X(\omega) = \sum_i x_i \mathbb{I}_{X = x_i}(\omega)$$

If we define

$$\varepsilon(x) = \begin{cases} 1 : x \geq 0 \\ 0 : x < 0 \end{cases}$$

Then

$$F(x) = \sum_i p_i \varepsilon(x - x_i)$$

Theorem

Theorem 14. *Let $\{p_k\}$ be a collection of non-negative real numbers such that $\sum_k p_k = 1$, then $\{p_k\}$ is the probability mass function (PMF) for some random variable X.*

The other type of random variable is those without jump points.

Definition

Definition 20. *Let X be a random variable defined on $(\Omega, \mathscr{F}, \mathbb{P})$ with distribution function F, then X is said to be of the continuous type if F is absolutely continuous, i.e there exists a non-negative function $f(x)$ such that*

$$\forall x \in \mathbb{R} : F(x) = \int_{-\infty}^{x} f(t)dt$$

The function f is called the probability density function (PDF) of the random variable X.

If F is absolutely continuous and f is continuous at x, then

$$F'(x) = \frac{dF(x)}{dx} = f(x)$$

Theorem

Theorem 15. *Let X be a random variable of the continuous type with probability density function f, then $\forall B \in \mathfrak{B}_{\mathbb{R}}$, we have*

$$\mathbb{P}(B) = \int_B f(t)dt$$

Theorem

Theorem 16. *Every non-negative real function f that is integrable over $\mathbb{R}$ and satisfies $\int_{-\infty}^{+\infty} f(x)dx = 1$ is the probability density function for some random variable X.*

Proof. We need to find such a F, where

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

Then by definition of $f(x)$, clearly $F(-\infty) = 0$, $F(+\infty) = 1$, and if $x_2 \geq x_1$, then

$$F(x_2) = \int_{-\infty}^{x_1} f(t)dt + \int_{x_1}^{x_2} f(t)dt \geq \int_{-\infty}^{x_1} f(t)dt = F(x_1)$$

So F is non-decreasing. Furthermore F is continuous (hence right-continuous), thus that finishes the proof. ■

Theorem

Theorem 17. *Let X be a random variable, then*

$$\mathbb{P}(x=a) := \lim_{t\to a^-} \mathbb{P}\{t < x \leq a\}$$

Now let's look at some examples:

Example 1. Suppose a random variable X is distributed according to $f_X(x)$, written as $X \sim f_X(x)$, satisfies

$$X \sim f_X(x) = \begin{cases} x : 0 < x \leq 1 \\ 2-x : 1 < x \leq 2 \\ 0 : \text{otherwise} \end{cases}$$

Find the distribution function F.

Solution: First we know that $\int_{\mathbb{R}} f_X(x)dx = 1$. By definition, $F(x)$ is given by

$$F(x) = \int_{-\infty}^{x} f(t)dt = \begin{cases} 0 : x \leq 0 \\ \\ \int_0^x tdt = \frac{1}{2}x^2 : 0 < x \leq 1 \\ \\ \int_0^1 tdt + \int_1^x (2-t)dt = 2x - \frac{1}{2}x^2 - 1 : 1 < x \leq 2 : 1 < x \leq 2 \\ \\ 1 : x > 2 \end{cases}$$

The graph of $F(x)$ is given by

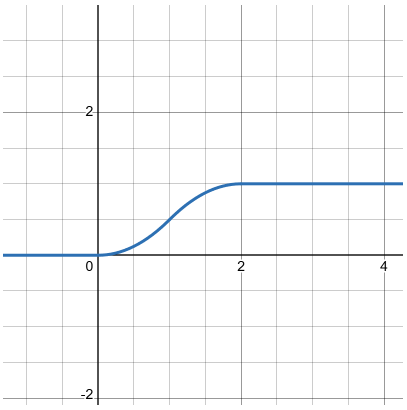


Figure 2.1: Graph of distribution function

Example 2. Suppose X is a random variable whose distribution function is given by

$$F(x) = \begin{cases} 0 : x < 0 \\ \frac{1}{2} : x = 0 \\ \frac{1}{2} + \frac{1}{2}x : 0 < x < 1 \\ 1 : x \geq 1 \end{cases}$$

Then we notice that there is a jump point at $x = 0$, so we may not differentiate directly, instead we write $F(x)$ as

$$F(x) = \frac{1}{2}F_d(x) + \frac{1}{2}F_c(x)$$

where

$$F_d(x) = \begin{cases} 0 : x < 0 \\ 1 : x \geq 0 \end{cases} \quad ; F_c(x) = \begin{cases} 0 : x \leq 0 \\ x : 0 < x < 1 \\ 1 : 1 \leq x \end{cases}$$

So in this case, $F_c(x)$ has no jump points, thus

$$f_c(x) = \frac{dF_c(x)}{dx} = \begin{cases} 1 : 0 < x < 1 \\ 0 : \text{otherwise} \end{cases}$$

and $F_d(x)$ is the distribution function of X degenerate at $x = 0$, i.e $F_d(x) = \mathbb{P}(x = 0) = 1$.

Functions of Random Variables

Theorem

Theorem 18. *Suppose X is a random variable, let g be a Borel measurable function on $\mathbb{R}$, then $Y = g(X)$ is also a random variable.*

Proof. We let $Y = g(X)$, then $F_Y(y) := \{g(X) \leq y\} = \{X \in g^{-1}((-\infty, y])\}$, which is measurable. ■

Example 1. Suppose X is a random variable with distribution function F_X, then what about $Y = |X|$?

Solution: We know that

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(|x| \leq y) = \mathbb{P}(-y \leq x \leq y)$$

If X is of the continuous type, then we have

$$F_Y(y) = F_X(y) - F_X(-y),$$

if X is of the discrete type, then we have

$$F_Y(y) = F_X(y) - F_X(-y^-), \text{since } X \text{ has a jump discontinuity at } -y.$$

Example 2. Suppose $X \sim N(0,1)$ where

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Where X satisfies normal distribution, which general form is

$$X \sim N(\mu, \sigma^2), f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

then what is F_Y, where $Y = |X|$?
Solution: As from above, we know that

$$F_Y(y) = F_X(y) - F_X(-y) = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx - \int_{-\infty}^{-y} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx.$$

Theorem

Theorem 19. *If X is a continuous random variable and g is differentiable, then $g(X)$ is also a continuous random variable.*

Back to example 2, we then have

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \frac{dF_X(y)}{dy} - \frac{dF_X(-y)}{dy} = f_x(y) + f_X(-y),$$

thus

$$f_Y(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} + \frac{1}{\sqrt{2\pi}} e^{-\frac{(-y)^2}{2}} = \sqrt{\frac{2}{\pi}} e^{-\frac{y^2}{2}}.$$

Example 3. Suppose $X \sim Poisson(\lambda)$, given by $\mathbb{P}(X = k) := \frac{e^{-\lambda}\lambda^k}{k!}$ where $k \in \mathbb{N}_0, \lambda > 0$. Now suppose $Y = X^2 + 3$, then

$$\mathbb{P}(Y = y) = \mathbb{P}(g(x) = y) = \mathbb{P}(x = \sqrt{y-3}) = \frac{e^{-\lambda}\lambda^{\sqrt{y-3}}}{(\sqrt{y-3})!}.$$

Theorem

Theorem 20. *Let X be a random variable of the continuous type, with probability density function f. Let $g(x)$ be differentiable, and $|g'(x)| > 0$ for all x, then $Y = g(X)$ is also a random variable of the continuous type with probability density function given by*

$$h_Y(y) = \begin{cases} f_X(g^{-1}(y)) \cdot \left| \frac{d}{dy} g^{-1}(y) \right| : \alpha < y < \beta \\ 0 : \textit{otherwise} \end{cases}$$

where $\alpha = \min\{g(-\infty), g(+\infty)\}; \beta = \max\{g(-\infty), g(+\infty)\}$.

Proof. Suppose $g'(x) > 0$, then g is continuous and strictly increasing, thus $\lim \alpha, \beta$ exists (possibly infinity), and its inverse $x = g^{-1}(y)$ also exists, differentiable, continuous and strictly increasing. The distribution function of Y for $\alpha < y < \beta$ is given by

$$\mathbb{P}(Y \leq y) = \mathbb{P}(X \leq g^{-1}(y))$$

The probability density function of g is obtained by differentiation:

$$\begin{aligned} h_Y(y) = \frac{d}{dy} F_Y(y) &= \frac{d}{dy} \mathbb{P}(Y \leq y) \\ &= \frac{d}{dy} \mathbb{P}(X \leq g^{-1}(y)) \\ &= \frac{d}{dy} F_X(g^{-1}(y)) \\ &= f_X(g^{-1}(y)) \cdot \frac{d}{dy} g^{-1}(y) \end{aligned}$$

Likewise, when $g'(x) < 0$, we have

$$h(y) = -f_X(g^{-1}(y)) \cdot \frac{d}{dy} g^{-1}(y)$$

And that completes the proof. ■

Example 1. Given that

$$X \sim f_X(x) = \begin{cases} \dfrac{2x}{\pi^2} : 0 < x < \pi \\ \\ 0 : \text{otherwise} \end{cases}$$

Let $Y = \sin(X)$, then find the probability density function of Y.

Solution: Let $Y = \sin(X) = g(X)$, then according to the figure below, we have

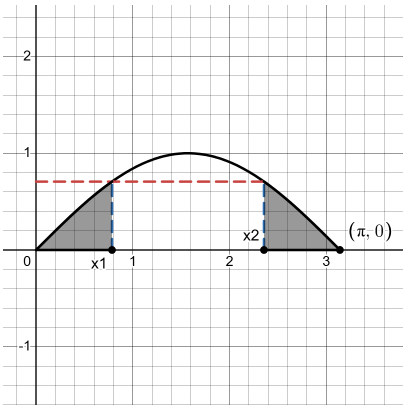


Figure 2.2: Solutions to $\sin(X) \leq y$

$\mathbb{P}(Y \leq y) = \mathbb{P}(\sin(x) \leq y) = \mathbb{P}(x \in (0,x_1) \cap x \in (x_2,\pi))$ where $x_1 = \sin^{-1}(y), x_2 = \pi - \sin^{-1}(y)$. Thus we have

$$\begin{aligned}\mathbb{P}(Y \leq y) &= \mathbb{P}(x \in (0,x_1)) + \mathbb{P}(x \in (x_2,\pi)) \\ &= \int_0^{x_1} f_X(x)dx + \int_{x_2}^{\pi} f_X(x)dx \\ &= \left(\frac{x_1}{\pi}\right)^2 + 1 - \left(\frac{x_2}{\pi}\right)^2\end{aligned}$$

So we have

$$h_Y(y) = \frac{d}{dy}F_Y(y) = \frac{d}{dy}\mathbb{P}(Y \leq y) = \frac{d}{dy}\left[\left(\frac{\sin^{-1}(x_1)}{\pi}\right)^2 + 1 - \left(\frac{\sin^{-1}(x_2)}{\pi}\right)^2\right]$$

and which simplifies to

$$h_Y(y) = \begin{cases} \dfrac{2}{\pi\sqrt{1-y^2}} : 0 < y < 1 \\ \\ 0 : \text{otherwise} \end{cases}.$$

Chapter 3

Moments and Generating Functions

Moments of a Distribution Function

Definition

Definition 21. *Let X be a random variable of the discrete type (continuous type) with probability mass function (probability density function) defined as $p_k := \mathbb{P}\{X = x_k\}, k = 1, 2, \cdots, (f_X)$, then if*

$$\sum |x_k| p_k < +\infty \quad \left(\int_{\mathbb{R}} |x| f_X(x) dx < +\infty \right),$$

We then say that the expected value of X exists, and denote

$$\mu_X = \mathbb{E}(X) = \sum x_k p_k \left(= \int_{\mathbb{R}} x f_X(x) dx \right)$$

Remark: Sometimes $\sum x_k p_k$ converges but not for $\sum |x_k| p_k$, in this case we say $\mathbb{E}(X)$ doesn't exist. The same idea applies for $\int_{\mathbb{R}} x f_X(x) dx$.

Corollary

Corollary 18. *Suppose X is a random variable and $\mathbb{E}(X)$ exists, then if $a, b \in \mathbb{R}$ and $\mathbb{E}(|aX + b|) < +\infty$, we then have $\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$*

Theorem

Theorem 21. *If X is a random variable of the discrete type, g is a Borel measurable function on $\mathbb{R}$, if $Y = g(X)$, then*

$$\mathbb{E}(Y) = \sum_k g(x_k)\mathbb{P}\{X = x_j\} = \sum_m y_m \mathbb{P}\{Y = y_m\}$$

On the other hand, if X is a random variable of the continuous type with probability density function f, and g is a Borel measurable function, then denote $Y = g(X)$ and $\mathbb{E}(|g(x)|) < +\infty$, where

$$\mathbb{E}(Y) = \int_{\mathbb{R}} g(x)f(x) = \int_{\mathbb{R}} yh(y)dy$$

Example 1. Suppose X has the uniform distribution on the first N natural numbers, i,e $\mathbb{P}(X = k) = \frac{1}{N}$ where $k = 1, 2, \cdots, N$, then

$$\mathbb{E}(X) = \sum_{k=1}^{N} k\frac{1}{N} = \frac{N+1}{2},$$

$$\mathbb{E}(X^2) = \sum_{k=1}^{N} k^2\frac{1}{N} = \frac{(N+1)(2N+1)}{6}.$$

Example 2. Let X be an random variable with probability density function defined by

$$f_X(x) = \begin{cases} \frac{2}{x^3} : x \geq 1 \\ \\ 0 : x < 1 \end{cases}.$$

Then

$$\mathbb{E}(X) = \int_{-\infty}^{1} x \cdot 0dx + \int_{1}^{\infty} x \cdot \frac{2}{x^3}dx = 2,$$

However

$$\mathbb{E}(X^2) = \int_1^{\infty} \frac{2}{x}dx \text{ diverges, thus doesn't exist.}$$

Theorem

Theorem 22. *Let X be a random variable, if $\mathbb{E}(|X|^t)$ exists for some $t > 0$, then for any $s \in (0, t)$, $\mathbb{E}(|X|^s)$ exists.*

Proof. We will prove for continuous random variable. Let f be the PDF of the random variable X, then

$$\begin{aligned} \mathbb{E}|X|^s &= \int_{|x|^s \leq 1} |x|^s f(x)dx + \int_{|x|^s > 1} |x|^s f(x)dx \\ &\leq \mathbb{P}\{|X|^s \leq 1\} + \mathbb{E}|X|^t \\ &< +\infty. \end{aligned}$$

A discrete random variable can be proven in a similar way. ■

Theorem

Theorem 23. *Let X be an random variable, if $\mathbb{E}(|X|^k)$ exists for some $k > 0$, then*

$$\lim_{n\to\infty} n^k \mathbb{P}\{|X| > n\} = 0$$

Proof. We will prove for the case when X is of the continuous type. A similar proof can be used for discrete type. Denote $f_X(x)$ to be the probability density function of X, then

$$\int_{-\infty}^{\infty} |x|^k f_X(x)dx = \lim_{n\to\infty} \int_{-n}^{n} |x|^k f_X(x)dx < +\infty$$

Then

$$\lim_{n\to\infty} \int_{|x|>n} |x|^k f_X(x)dx \geq n^k \mathbb{P}\{|X| > n\} = 0$$

■

Probabilities of the type $\mathbb{P}\{|X| > n\}$ or either of its components are called tail probabilities.

Remark: The converse of theorem 24 is not necessarily true.

Theorem

Theorem 24. *Let X be a non-negative random variable with distribution function $F_X(x)$, then*

$$\mathbb{E}(X) = \int_0^{\infty} [1 - F_X(x)]dx$$

Proof. Since X is non-negative, then

$$\mathbb{E}(X) = \int_0^{+\infty} x f_X(x)dx = \lim_{n\to\infty} \int_0^n x f_X(x)dx$$

According to integration by parts, we have

$$\begin{aligned}\int_0^n x f_X(x) &= xF_X(x)\Big|_0^n - \int_0^n F_X(x)dx \\ &= -n[1 - F_X(n)] + \int_0^n [1 - F_X(x)]dx\end{aligned}$$

As $n \to \infty$, $-n[1 - F_X(n)] \to 0$ and hence we finished the proof. ■

Corollary

Corollary 19. *For any random variable X, $\mathbb{E}(X) < +\infty$ if and only if both $\int_{-\infty}^{0} \mathbb{P}\{X \leq x\}dx$ and $\int_0^{\infty} \mathbb{P}\{X \leq x\}dx$ converge.*

Corollary

Corollary 20. *Let X be a random variable, $\alpha > 0$, then*

$$\mathbb{E}(|X|^{\alpha}) < +\infty \Longleftrightarrow \sum_{n=1}^{\infty} \mathbb{P}\{|X| > n^{\frac{1}{\alpha}}\} < +\infty$$

Theorem

Theorem 25. *Let X be a random variable satisfying $\lim_{n\to\infty} n^{\alpha}\mathbb{P}\{|x| > n\} = 0$ with $\alpha > 0$, then for any $0 < \beta < \alpha$, we have $\mathbb{E}(|X|^{\beta}) < +\infty$.*

Proof. Given any $\varepsilon > 0$, we can always choose a large enough N such that $\forall n \geq N$,

$$\mathbb{P}\{|x| > n\} < \frac{\varepsilon}{n^{\alpha}},$$

and

$$\mathbb{E}(|X|^{\beta}) = \beta \int_0^N x^{\beta-1}\mathbb{P}\{|X| > x\}dx + \beta \int_N^{\infty} x^{\beta-1}\mathbb{P}\{|X| > x\}dx \qquad (*)$$

By theorem 25, we know that

$$\begin{aligned}
\mathbb{E}|X|^{\beta} &= \int_0^{+\infty} \mathbb{P}\{|X|^{\beta} > x\}dx \\
&= \int_0^{\infty} \mathbb{P}\{|X| > x^{\frac{1}{\beta}}\}dx \\
&= \beta \int_0^{+\infty} u^{\beta-1}\mathbb{P}\{|X| > u\}du \ \text{(denote } x^{\frac{1}{\beta}} = u\text{).} \quad (**)
\end{aligned}$$

Now combine $(*), (**)$, we have

$$\begin{aligned}
\mathbb{E}|X|^{\beta} &\leq \beta \int_0^N x^{\beta-1}dx + \beta \int_N^{\infty} x^{\beta-1}\mathbb{P}\{|X| > x\}dx \\
&\leq N^{\beta} + \beta\varepsilon \int_N^{\infty} x^{\beta-\alpha-1}dx < +\infty.
\end{aligned}$$

■

Definition

Definition 22. *Let k be a positive integer and c be a constant, if $\mathbb{E}(X-c)^k$ exists, we then call it the moment of order k about the point c. If we take $c = \mathbb{E}(X) = \mu_X$, then we call $E(X-\mu_X)^k$ to be the central moment of order k about the mean. Generally $\mathbb{E}(X^n)$ is the nth moment of X and $\mathbb{E}|X|^{\alpha}$ is the αth absolutely moment of X.*

We will mainly study one special case, that is when $k = 2$:

Definition

Definition 23. *Let X be a random variable, if $\mathbb{E}(X^2)$ exists, we denote $\mathbb{E}(X - \mathbb{E}(X))^2$ to be the variance of X, written as $\sigma_X^2 = \mathbf{Var}(X) = \mathbb{E}(x-\mu_X)^2$.*

Remark: We know that

$$\begin{aligned}\mathbf{Var}(X) &= \mathbb{E}[(x-\mu_x)^2] \\ &= \mathbb{E}[X^2 - 2X\mu_X + \mu_X^2] \\ &= \mathbb{E}(x^2) - 2\mu_X\mathbb{E}(X) + \mu_X^2 \\ &= \mathbb{E}(X^2) - (\mathbb{E}(X))^2\end{aligned}$$

Corollary

Corollary 21. *Let X be a random variable, $a, b \in \mathbb{R}$. Then*

$$\mathbf{Var}(aX+b) = a^2\,\mathbf{Var}(X)$$

Definition

Definition 24. *Let X be a random variable and $\mathbb{E}|X|^2 < +\infty$, then define*

$$Z = \frac{X - \mathbb{E}(X)}{\sqrt{\mathbf{Var}(X)}} = \frac{X-\mu_X}{\sigma_X}$$

to be the standardized random variable, and $\mathbb{E}(Z) = 0, \mathbf{Var}(Z) = 1$.

Generating Functions

Here, we will introduce and discuss moment generating functions (MGFs). Moment generating functions are useful for several reasons, one of which is their application to analysis of sums of random variables. Before discussing MGFs, let's define moments.

Definition

Definition 25. *The nth moment of a random variable X is dedfined to be $\mathbb{E}[X^n]$. The nth central moment of X is defined to be $\mathbb{E}[(X - \mathbb{E}X)^n]$.*

Definition

Definition 26. *The function $P(s) = \sum_{k=0}^{\infty} p_k s^k$ where $\{p_k\}$ is the probability mass function of a discrete random variable X, is called the probability generating function (PGF). It's clear that $P(s)$ exists for $|s| < 1$*

Corollary

Corollary 22. *For the probability generating function $P(s)$ of the discrete type, we have*

$$\frac{1}{n!}P^{(n)}(s)\bigg|_{s=0} = p_n = \mathbb{P}\{X = n\}.$$

Also, in $P(s)$, we notice that

$$P'(s)\bigg|_{s=1} = \mathbb{E}(X); P''(s)\bigg|_{s=1} = \mathbb{E}[X(X-1)],$$

and

$$\mathbb{E}(X^2) = \mathbb{E}[X(X-1)] + \mathbb{E}(X).$$

Example: Consider $X \sim Possion(\lambda)$, where

$$\mathbb{P}(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

Then we know that

$$P(s) = \sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k}{k!}s^k = e^{-\lambda}\sum_{k=0}^{\infty}\frac{(\lambda s)^k}{k!} = e^{-\lambda(1-s)}$$

Thus

$$\mathbb{E}(X) = P'(s)\bigg|_{s=1} = \lambda e^{-\lambda(1-s)}\bigg|_{s=1} = \lambda,$$

and

$$\mathbb{E}[X(X-1)] = P''(s)\Big|_{s=1} = \lambda^2 e^{-\lambda(1-s)}\Big|_{s=1} = \lambda^2$$

Thus we know that

$$\mathbb{E}(X^2) = \mathbb{E}(X) + \mathbb{E}[X(1-X)] = \lambda^2 + \lambda,$$

and

$$\mathbf{Var}(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 = \lambda.$$

Definition

Definition 27. *Let X be a random variable, the function*

$$M_X(s) = \mathbb{E}[e^{sX}] = \int_{\mathbb{R}} e^{sx} f_X(x)dx$$

is known as the moment generating function (MGF) of the continuous type of random variable X, if $\mathbb{E}[e^{SX}]$ exists in some neighborhood of the origin. For discrete type, we have

$$M_X(s) = \mathbb{E}[e^{sX}] = \sum e^{sx_i} p_i$$

Example: For each of the following random variables, find the MGF:

(a) X is a discrete random variable with pmf $P_X(k) = \begin{cases} \frac{1}{3} : k = 1 \\ \frac{2}{3} : k = 2 \end{cases}$;

(b) Y is a $Uniform(0,1)$ random variable.

Solution:
(a) We have

$$M_X(s) = \mathbb{E}[e^{sX}] = \frac{1}{3}e^s + \frac{2}{3}e^{2s}.$$

(b) We have

$$M_Y(s) = \mathbb{E}[e^{sY}] = \int_0^1 e^{sy}dy = \frac{e^s - 1}{s}.$$

Example: Suppose $X \sim \exp(\lambda), \lambda > 0$ which is given by

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} : x > 0 \\ 0 : \text{otherwise} \end{cases}$$

In this case

$$\begin{aligned} \mathbb{E}[e^{sx}] &= \int_{\mathbb{R}} e^{sx} f_X(x)dx \\ &= \int_0^{+\infty} e^{sx}\lambda e^{-\lambda x}dx = \lambda \left[-\frac{e^{-(\lambda - s)x}}{\lambda - s} \right]_{x=0}^{x=+\infty}, \end{aligned}$$

and

$$M_X(s) = \begin{cases} \dfrac{\lambda}{\lambda - s} : s < \lambda \\ \\ \text{DOES NOT EXIST} \ : s \geq \lambda \end{cases} .$$

We will introduce two theorems, however we will omit the proof:

Theorem

Theorem 26. *The moment generating function uniquely determines a distribution function and conversely, if the moment generating function exists, then it is unique, i.e the map*

$$\Gamma : f_X \longrightarrow m(s) := \int_{\mathbb{R}} e^{sx} f_X(x) dx$$

is bijective.

Theorem

Theorem 27. *If the moment generating function $M_X(s)$ of a random variable X exists for some $|s| < s_0$, then the derivatives of all order exist at $s = 0$, and*

$$M_X^{(k)}(s)\Big|_{s=0} = \mathbb{E}(X^k)$$

for positive integers k.

Why is the MGF useful? There are basically two reasons for this. First, the MGF of X gives us all moments of X . That is why it is called the moment generating function. Second, the MGF (if it exists) uniquely determines the distribution. That is, if two random variables have the same MGF, then they must have the same distribution. Thus, if you find the MGF of a random variable, you have indeed determined its distribution. We will see that this method is very useful when we work on sums of several independent random variables. Let's discuss these in detail.

Recall that

$$e^x = \sum_{k=0}^{+\infty} \frac{x^k}{k!}$$

where we have

$$e^{sX} = \sum_{k=0}^{+\infty} \frac{(sX)^k}{k!} = \sum_{k=0}^{+\infty} \frac{X^k s^k}{k!}$$

Thus we have

$$M_X(s) = \mathbb{E}[e^{sX}] = \sum_{k=0}^{+\infty} \mathbb{E}[X^k] \frac{s^k}{k!}$$

So, we conclude that the kth moment of X is the coefficient of $\frac{s^k}{k!}$ in the Taylor series of $M_X(s)$. If we have the Taylor series of $M_X(s)$, we can then obtain all moments of X.

Example: If $Y \sim Uniform(0,1)$, find $\mathbb{E}[Y^k]$.

Solution: Recall that

$$M_Y(s) = \frac{e^s - 1}{s}$$

Also

$$\begin{aligned} M_Y(s) &= \frac{1}{s}\left(\sum_{k=0}^{+\infty} \frac{s^k}{k!} - 1\right) \\ &= \frac{1}{s}\sum_{k=1}^{+\infty} \frac{s^k}{k!} \\ &= \sum_{k=1}^{+\infty} \frac{s^{k-1}}{k!} \\ &= \sum_{k=0}^{+\infty} \frac{1}{k+1}\frac{s^k}{k!}. \end{aligned}$$

Thus the coefficient of $\frac{s^k}{k!}$ in the Taylor series for $M_Y(s)$ is $\frac{1}{k+1}$, thus

$$\mathbb{E}[X^k] = \frac{1}{k+1}.$$

Corollary

Corollary 23. *We can also obtain all moments of X^k from its MGF:*

$$M_X(s) = \sum_{k=0}^{+\infty} \mathbb{E}[X^k]\frac{s^k}{k!},$$

$$\mathbb{E}[X^k] = \frac{d^k}{ds^k} M_X(s)\Big|_{s=0}.$$

Example: Let $X \sim Exponential(\lambda)$, find all of its moments $\mathbb{E}[X^k]$.

Solution: Recall that

$$M_X(s) = \frac{\lambda}{\lambda - s}, s < \lambda$$

So

$$\begin{aligned} M_X(s) &= \frac{\lambda}{\lambda - s} \\ &= \frac{1}{1 - \frac{s}{\lambda}} \\ &= \sum_{k=0}^{+\infty} \left(\frac{s}{\lambda}\right)^j, \left|\frac{s}{\lambda}\right| < 1 \\ &= \sum_{k=0}^{+\infty} \frac{k!}{\lambda^k}\frac{s^k}{k!}. \end{aligned}$$

And thus we conclude that

$$\mathbb{E}[X^k] = \frac{k!}{\lambda^k}.$$

Example: Suppose $X \sim Geometric(p)$ which is given by

$$p_k = \mathbb{P}(X = k) = p(1-p)^k,$$

then

$$\begin{aligned} M_X(s) &= \mathbb{E}(e^{SX}) \\ &= \sum_{n=0}^{\infty} e^{sn}\mathbb{P}(X = n) \\ &= p\sum_{n=0}^{\infty}[(1-p)e^s]^n \\ &= p\frac{1}{1-(1-p)e^s}, \text{ if } s < -\ln(1-p). \end{aligned}$$

Moments Inequalities

We shall discuss several moments inequalities in this section.

Theorem

Theorem 28. *(Markov's Inequality)*
Let $h(x)$ be a non-negative function of a random variable X, if $\mathbb{E}(X)$ exists, then for every $\varepsilon > 0$, we have

$$\mathbb{P}\{h(x) \geq \varepsilon\} \leq \frac{\mathbb{E}[h(x)]}{\varepsilon}$$

Proof. Define $A := \{x : h(x) \geq \varepsilon\}$ then define $f_X(x)$ to be the probability density function, then we have

$$\begin{aligned}
\mathbb{E}[h(x)] &= \int_{\mathbb{R}} h(x) f_X(x) dx \\
&= \int_A h(x) f_X(x) dx + \int_{A^C} h(x) f_X(x) dx \\
&\geq \int_A \varepsilon f_X(x) dx + \int_{A^C} h(x) f_X(x) dx \\
&\geq \varepsilon \mathbb{P}\{h(x) \geq \varepsilon\}
\end{aligned}$$

■

Corollary

Corollary 24. *Let $h(x) = |x|^r, \varepsilon = k^r$ where $r > 0, k > 0$, then*

$$\mathbb{P}\{|X| \geq k\} \leq \frac{\mathbb{E}|X|^r}{k^r}$$

Corollary

Corollary 25. *Let $h(x) = (x - \mu_X)^2, \varepsilon = k^2\sigma_X^2$ where $\mu_X = \mathbb{E}(X)$ and $\sigma_X^2 = \mathbf{Var}(X)$, then*

$$\mathbb{P}\{|X - \mu_X| \geq k\sigma_X\} \leq \frac{1}{k^2}$$

In corollary 12, if we choose $k = 3$, then we get

$$\mathbb{P}\{|X - \mu_X| \geq 3\sigma_X\} \leq \frac{1}{9} \approx 11\%$$

Corollary 12 will also result in another inequality:

Theorem

Theorem 29. *(Chebyshev's Inequality)*
If X is any random variable, then for any $b > 0$ we have

$$\mathbb{P}\{|X - \mathbb{E}(X)| \geq b\} \leq \frac{\mathbf{Var}(X)}{b^2}$$

Proof. According to Markov's Inequality, we have

$$\mathbb{P}\{(X - \mathbb{E}(X))^2 \geq b^2\} \leq \frac{\mathbb{E}[(X - \mathbb{E}(X))^2]}{b^2}.$$

■

Here are some extensions of Chebyshev's Inequality, namely Gauss Inequality and Vysochanskij–Petunin Inequality.

Corollary

Corollary 26. *(Gauss Inequality)*
Assume $X \sim f$ where f is uni modal (meaning that f obtain a single maximum) with modal ν where

$$\nu = \arg\max_x f(x)$$

and define $\tau^2 = \mathbb{E}(X - \nu)^2$, then we have

$$\mathbb{P}_r\{|X - \nu| \geq \varepsilon\} \leq \begin{cases} \frac{4\tau^2}{9\varepsilon^2} : \varepsilon \geq \frac{2}{\sqrt{3}}\tau \\ 1 - \frac{\varepsilon}{\tau\sqrt{3}} : \varepsilon < \frac{2}{\sqrt{3}}\tau \end{cases}$$

Corollary

Corollary 27. *(Vysochanskij–Petunin Inequality)*
Let $X \sim f$ where f is uni modal and define $\xi^2 = \mathbb{E}[(x - \alpha)^2]$ for any $\alpha \in \mathbb{R}$, then $\forall \varepsilon > 0$, we have

$$\mathbb{P}_r\{|X - \alpha| \geq \varepsilon\} \leq \begin{cases} \frac{4\xi^2}{9\varepsilon^2} : \varepsilon \geq \sqrt{\frac{8}{3}}\xi \\ \frac{4\xi^2}{3\varepsilon^2} - \frac{1}{3} : \varepsilon < \sqrt{\frac{8}{3}}\xi \end{cases}$$

The proof for corollary 13 and 14 will not be proved, since those require some work.

Theorem

Theorem 30. *(Chernoff Bounds)*
Let X be a random variable, then $\forall a \in \mathbb{R}$,

$$\mathbb{P}\{X \geq a\} \leq e^{-sa} M_X(s) : s > 0$$

and

$$\mathbb{P}\{X \geq a\} \geq e^{-sa} M_X(s) : s < 0$$

Where $M_X(s)$ denotes the moment generating function of the random variable X.

Proof. For $t > 0$, we have

$$\mathbb{P}\{X \geq a\} = \mathbb{P}\{e^{tX} \geq e^{ta}\} \leq \mathbb{E}(e^{tX})e^{-ta}, \text{by Markov's Inequality.}$$

A similar proof can be done when $t < 0$. ■

Corollary

Corollary 28. *Let Z to be the standard normal random variable, then for $a > 0$, we have*

$$\mathbb{P}\{Z \geq a\} \leq e^{-a^2/2}$$

and similarly, for $a < 0$, we have

$$\mathbb{P}\{Z \leq a\} \leq e^{-a^2/2}$$

Corollary

Corollary 29. *Let X be a random variable with $\mathbb{E}(X) = 0$ and $\mathbf{Var}(X) = \sigma_X^2$, then*

$$\mathbb{P}(X \geq x) \leq \frac{\sigma_X^2}{\sigma_X^2 + x^2} : x > 0$$

and

$$\mathbb{P}(X \geq x) \geq \frac{\sigma_X^2}{\sigma_X^2 + x^2} : x < 0$$

Proof. Proof is left as an exercise :) ■

Theorem

Theorem 31. *(Minkouski Inequality)*
Let X, Y be random variables, $r > 0$. If $\mathbb{E}X^r, \mathbb{E}Y^r$ exist, then so is $\mathbb{E}|X+Y|^r$ and

$$\mathbb{E}|X+Y|^r \leq C_r[\mathbb{E}|X|^r + \mathbb{E}|Y|^r]$$

where $C_r = \begin{cases} 1 : 0 \leq r \leq 1 \\ 2^{r-1} : r > 1 \end{cases}$.

Discrete Distributions

3.4.1 Two-Point Distribution (Bernoulli Distribution)

Definition

Definition 28. *We say that an random variable X has a two-point distribution if it takes only two values x_1 and x_2, with probabilities*

$$\mathbb{P}\{X = x_1\} = p \ \ , \mathbb{P}\{X = x_2\} = 1 - p, \ \ 0 < p < 1$$

We may also write

$$X = x_1 \mathbf{1}_{X=X_1} + x_2 \mathbf{1}_{X=x_2}$$

The probability mass function is given by

$$f(x) = \begin{cases} p, x = x_1 \\ 1 - p, x = x_2 \\ 0, \text{otherwise} \end{cases}$$

and the cumulated distribution function is given by (where we assume $x_1 < x_2$)

$$F(x) = \begin{cases} 0, x < x_1 \\ p, x_1 \leq x < x_2 \\ 1, x \geq x_2 \end{cases}$$

The expected value is given by

$$\mathbb{E}(X) = px_1 + (1 - p)x_2,$$

and in general,

$$\mathbb{E}(X^k) = px_1^k + (1 - p)x_2^k.$$

The variance is given by

$$\mathbf{Var}(X) = p(1 - p)(x_1 - x_2)^2.$$

Lastly, the generating function is given by

$$M(t) = pe^{tx_1} + (1 - p)e^{tx_2}, t \in \mathbb{R}.$$

Now we have a special case for two-point distribution, if we have $x_1 = 1, x_2 = 0$, we get the Bernoulli random variable:

$$\mathbb{P}\{X = 1\} = p, \mathbb{P}\{X = 0\} = 1 - p, \ \ 0 < p < 1$$

In Bernoulli random variable, we always assign p to be the probability of success and $1 - p$ to be the probability of failure.

Example 1: In a sequence of n Bernoulli trails with constant probability p of success (S) and $1 - p$ of failure (F), let Y_n be the number of times that the ordered combination SF occur, find its expected value

and variance.

Solution: Let Y_n denote the number of times the combination SF occur, and we define

$$f(X_i,X_{i+1}) = \begin{cases} 1, \text{if } X_i = S, X_{i+1} = F \\ 0, \text{otherwise} \end{cases}$$

where $i = 1,2,\cdots,n_1$, then we have

$$Y_n = \sum_{i=1}^{n-1} f(X_i,X_{i+1})$$

so

$$\mathbb{E}(Y_n) = (n-1)p(1-p)$$

and

$$\begin{aligned} \mathbb{E}(Y_n^2) &= \mathbb{E}\left[\sum_{i=1}^{n-1} f^2(X_i,X_{i+1})\right] + \mathbb{E}\left[\sum\sum_{i\neq j} f(X_i,X_{i+1})f(X_j,X_{j+1})\right] \\ &= (n-1)p(1-p) + (n-2)(n-3)p^2(1-p)^2. \end{aligned}$$

So $\mathbf{Var}(Y_n) = p(1-p)[n-1+p(1-p)(5-3n)]$.

3.4.2 Uniform Distribution on n Points

Definition

Definition 29. *X is said to have a uniform distribution on n points $\{x_1,x_2,\cdots,x_n\}$ if its probability mass function is of the form*

$$\mathbb{P}\{X = x_i\} = \frac{1}{n}, \quad i = 1,2,\cdots,n$$

So we have

$$\mathbb{E}(X) = \frac{1}{n}\sum_{i=1}^{n} x_i, E(X^k) = \frac{1}{n}\sum_{i=1}^{n} x_i^k$$

$$\mathbf{Var}(X) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

where $\bar{x}$ is the average of $x_1,x_2,\cdots,x_n$, given by

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The generating function is also given by

$$M(t) = \frac{1}{n}\sum_{i=1}^{n} e^{tx_i}, \ \ t \in \mathbb{R}$$

In a special case, if we let $x_i = i$, $i = 1, 2, \cdots, n$, we then have

$$\mathbb{E}(X) = \frac{n+1}{2}, E(X^2) = \frac{(n+1)(2n+1)}{6}, \mathbf{Var}(X) = \frac{n^2-1}{12}.$$

Example 2: A box contains tickets numbered 1 to N, let X be the largest number drawn in n random drawings with replacement, then we know that those tickets are uniformly distributed, and

$$\mathbb{P}\{X \leq k\} = \left(\frac{k}{N}\right)^n$$

So

$$\begin{aligned}\mathbb{P}\{X = k\} &= \mathbb{P}\{X \leq k\} - \mathbb{P}\{X \leq k-1\} \\ &= \left(\frac{k}{n}\right)^n - \left(\frac{k-1}{n}\right)^n.\end{aligned}$$

Also

$$\begin{aligned}\mathbb{E}(X) &= N^{-n} \sum_{k=1}^{N} k^{n+1} - (k-1)^{n+1} - (k-1)^n \\ &= N^{-n} \left[N^{n+1} - \sum_{k=1}^{N} (k-1)^n \right].\end{aligned}$$

3.4.3 Binomial Distribution

Binomial distribution can be viewed as the sum of n Bernoulli random variables with the same success probability p.

Definition

Definition 30. *We say that X has a binomial distribution with parameter p if its probability mass function is given by*

$$p_k = \mathbb{P}\{X = k\} = \binom{n}{k} p^k (1-p)^{n-k}$$

where $k = 0, 1, \cdots, n; 0 \leq p \leq 1$.

In Binomial distribution, we have

$$\mathbb{E}(X) = np, E(X^2) = n(n-1)p^2 + np, \mathbf{Var}(X) = np(1-p)$$

and most importantly,

$$\begin{aligned}M(t) &= \sum_{k=0}^{n} e^{tk} \binom{n}{k} p^k (1-p)^{n-k} \\ &= (1 - p + pe^t)^n, \ \ t \in \mathbb{R}.\end{aligned}$$

Example: Five fair coins are flipped. If the outcomes are assumed independent, find the probability mass function of the number of heads obtained.

Solution: If we denote X to be the number of heads, then X is a binomial random variable with parameters $n = 5, p = \frac{1}{2}$. Hence we have

$$\mathbb{P}\{X = 0\} = \binom{5}{0}\left(\frac{1}{2}\right)^0\left(1-\frac{1}{2}\right)^5 = \frac{1}{32}$$
$$\mathbb{P}\{X = 1\} = \binom{5}{1}\left(\frac{1}{2}\right)^1\left(1-\frac{1}{2}\right)^4 = \frac{5}{32}$$
$$\mathbb{P}\{X = 2\} = \binom{5}{2}\left(\frac{1}{2}\right)^2\left(1-\frac{1}{2}\right)^3 = \frac{10}{32}$$
$$\mathbb{P}\{X = 3\} = \binom{5}{3}\left(\frac{1}{2}\right)^3\left(1-\frac{1}{2}\right)^2 = \frac{10}{32}$$
$$\mathbb{P}\{X = 4\} = \binom{5}{4}\left(\frac{1}{2}\right)^4\left(1-\frac{1}{2}\right)^1 = \frac{5}{32}$$
$$\mathbb{P}\{X = 5\} = \binom{5}{5}\left(\frac{1}{2}\right)^5\left(1-\frac{1}{2}\right)^0 = \frac{1}{32}$$

Example: A communication system consists of n components, each of which will independently function with probability p. The total system will be able to operate effectively if at least one-half of its components function. For what values of p is a 5-component system better than a 3-component system?

Solution: The number of functioning components X is a binomial random variable with parameters n, p, so the probability that a 5-component system will be effective is given by

$$\binom{5}{3}p^3(1-p)^2 + \binom{5}{4}p^4(1-p)^1 + p^5$$

and similarly, for a 3-component system, the probability is given by

$$\binom{3}{2}p^2(1-p) + p^3$$

Hence, the 5-component system is effective if

$$\binom{5}{3}p^3(1-p)^2 + \binom{5}{4}p^4(1-p)^1 + p^5 > \binom{3}{2}p^2(1-p) + p^3$$

Which is,

$$p > \frac{1}{2}.$$

In general, when is a $(2k+1)$-component system better than a $(2k-1)$- component system?

3.4.4 Poisson Random Variable

If n independent trials, each of which results in a success with probability p, are performed, then, when n is large and p is small enough to make np moderate, the number of successes occurring is approximately a Poisson random variable with parameter $\lambda = np$.

The Poisson random variable has a tremendous range of applications in diverse areas because it may be used as an approximation for a binomial random variable with parameters (n, p) when n is large and p is small enough so that np is of moderate size. To see this, suppose that X is a binomial random variable with parameters (n, p), and let $\lambda = np$. Then

$$\begin{aligned}\mathbb{P}\{X = i\} &= \binom{n}{i} p^i (1-p)^{n-i} \\ &= \frac{n!}{(n-i)!i!} \left(\frac{\lambda}{n}\right)^i \left(1 - \frac{\lambda}{n}\right)^{n-i} \\ &= \frac{n(n-1)\cdots(n-i+1)}{n^i} \frac{\lambda^i}{i!} \frac{(1-\lambda/n)^n}{(1-\lambda/n)^i}\end{aligned}$$

Now, for large n and λ moderate, we have

$$\left(1 - \frac{\lambda}{n}\right)^n \approx e^{-\lambda} \quad \frac{n(n-1)\cdots(n-i+1)}{n^i} \approx 1 \quad \left(1 - \frac{\lambda}{n}\right)^i \approx 1$$

Hence we have

$$\mathbb{P}\{X = i\} \approx e^{-\lambda} \frac{\lambda^i}{i!}.$$

Definition

Definition 31. *A random variable X that takes one of the values $0, 1, 2, \cdots$ is said to be a Poisson random variable with parameter λ, if for some $\lambda > 0$,*

$$\mathbb{P}\{X = i\} = e^{-\lambda} \frac{\lambda^i}{i!}$$

We have the following properties of a Poisson variable:

$$\mathbb{E}(X) = \lambda; \mathbb{E}(X^2) = \lambda(\lambda + 1); \mathbf{Var}(X) = \lambda$$

Continuous Distribution

3.5.1 Uniform Distribution

Definition

Definition 32. *A random variable X is said to have a uniform distribution on the interval $[a,b]$, if its probability density function is given by*

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

And we denote by $X \sim Uniform[a,b]$.

Based on that, we may derive the distribution function $F(x)$ to be

$$F(x) = \begin{cases} 0 : x < a \\ \frac{x-a}{b-a} : a \leq x \leq b \\ 1 : x > b \end{cases}$$

Corollary

Corollary 30. *Suppose $X \sim U[a,b]$, then for all positive integer k,*

$$\mathbb{E}X = \frac{a+b}{2}; EX^k = \frac{b^{k+1} - a^{k+1}}{(k+1)(b-a)}$$

and

$$\mathbf{Var}(X) = \frac{(b-a)^2}{12}.$$

Proof. Left as an exercise to the reader. ■

Furthermore, we can derive the moment generating function of $X \sim Uniform[a,b]$, we have

$$\begin{aligned} M_X(s) = \mathbb{E}[e^{sX}] &= \int_a^b e^{sx} \frac{1}{b-a} dx \\ &= \frac{e^{sb} - e^{sa}}{s(b-a)}. \end{aligned}$$

3.5.2 Exponential Distribution

The exponential distribution is one of the widely used continuous distributions. It is often used to model the time elapsed between events. We will now mathematically define the exponential distribution, and

derive its mean and expected value. Then we will develop the intuition for the distribution and discuss several interesting properties that it has.

Definition

Definition 33. *A continuous random variable X is said to have an exponential distribution with parameter $\lambda > 0$, shown as $X \sim Exponential(\lambda)$, if its PDF is given by*

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ \\ 0 & otherwise \end{cases}$$

Below is a figure showing the graph of $f_X(x)$ with different parameter λ.

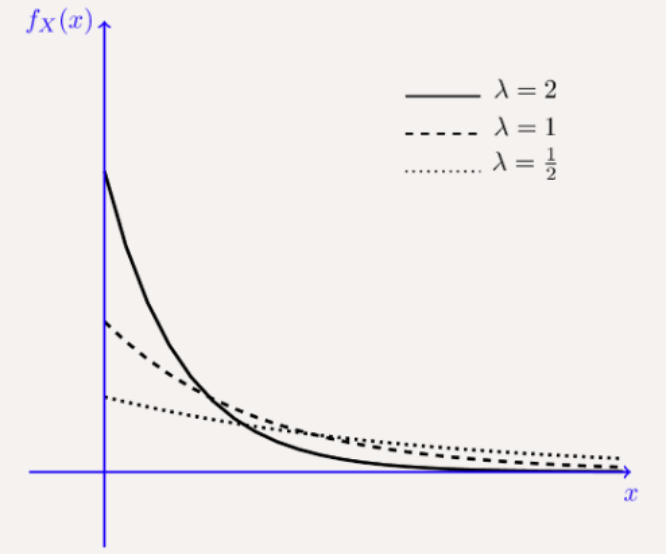


Figure 3.1: The exponential distribution with different parameter

We may also derive the distribution function of $X \sim Exponential(\lambda)$:

$$F_X(x) = \mathbb{P}\{X \leq x\} = \int_0^x \lambda e^{-\lambda t} dt = 1 - e^{-\lambda x}.$$

We can also find its expected value using integration by parts:

$$\begin{aligned} \mathbb{E}X &= \int_0^{+\infty} x\lambda e^{-\lambda x} dx \\ &= \frac{1}{\lambda} \int_0^{+\infty} y e^{-y} dy && \text{choosing } y = \lambda x \\ &= \frac{1}{\lambda} \left[-e^{-y} - y e^{-y} \right]_0^{+\infty} \\ &= \frac{1}{\lambda}. \end{aligned}$$

Now let's find $\mathbf{Var}(X)$. Since

$$\mathbb{E}X^2 = \int_0^{+\infty} x^2 \lambda e^{-\lambda x} dx = \frac{2}{\lambda^2}.$$

Thus we obtain

$$\mathbf{Var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \frac{1}{\lambda^2}.$$

Corollary

Corollary 31. *If $X \sim Exponential(\lambda)$, then*

$$\mathbb{E}X = \frac{1}{\lambda}, \mathbf{Var}(X) = \frac{1}{\lambda^2}.$$

Proof. This follows trivially from the arguments above. ■

To get some intuition for this interpretation of the exponential distribution, suppose you are waiting for an event to happen. For example, you are at a store and are waiting for the next customer. In each millisecond, the probability that a new customer enters the store is very small. You can imagine that, in each millisecond, a coin (with a very small $\mathbb{P}(H)$) is tossed, and if it lands heads a new customers enters. If you toss a coin every millisecond, the time until a new customer arrives approximately follows an exponential distribution.

The above interpretation of the exponential is useful in better understanding the properties of the exponential distribution. The most important of these properties is that the exponential distribution is memoryless. To see this, think of an exponential random variable in the sense of tossing a lot of coins until observing the first heads. If we toss the coin several times and do not observe a heads, from now on it is like we start all over again. In other words, the failed coin tosses do not impact the distribution of waiting time from now on. The reason for this is that the coin tosses are independent. We can state this formally as follows:

$$\mathbb{P}(X > x+a | X > a) = \mathbb{P}(X > x).$$

To see why this is true, we have

$$\begin{aligned} P(X > x+a | X > a) &= \frac{\mathbb{P}(X > x+a, X > a)}{\mathbb{P}(X > a)} \\ &= \frac{\mathbb{P}(X > x+a)}{\mathbb{P}(X > a)} \\ &= \frac{1 - F_X(x+a)}{1 - F_X(a)} \\ &= \frac{e^{-\lambda(x+a)}}{e^{-\lambda a}} \\ &= e^{-\lambda x} \\ &= \mathbb{P}(X \geq x). \end{aligned}$$

3.5.3 Normal (Gaussian) Distribution

The normal distribution is by far the most important probability distribution. One of the main reasons for that is the Central Limit Theorem (CLT) that we will discuss later. To give you an idea, the CLT states that if you add a large number of random variables, the distribution of the sum will be approximately normal under certain conditions. The importance of this result comes from the fact that many random variables in real life can be expressed as the sum of a large number of random variables and, by the CLT, we can argue that distribution of the sum should be normal. The CLT is one of the most important results in probability and we will discuss it later on. Here, we will introduce normal random variables.
We first define the standard normal random variable. We will then see that we can obtain other normal random variables by scaling and shifting a standard normal random variable.

Definition

Definition 34. *A continuous random variable Z is said to be a standard normal (standard Gaussian) random variable, shown as $Z \sim N(0,1)$, if its PDF is given by*

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{z^2}{2}\right\}, \qquad \text{for all } z \in \mathbb{R}$$

The PDF is a bell curve, given by the figure below:

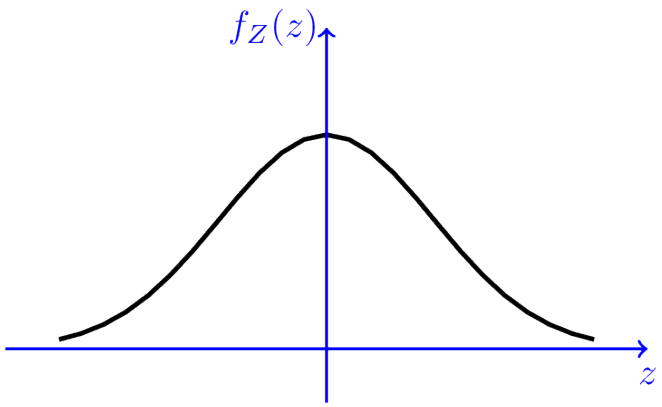


Figure 3.2: Standard Normal Distribution

And the total area under the curve is 1. This is because of the fact that

$$\int_{-\infty}^{+\infty} e^{-x^2} dx = \sqrt{\pi}.$$

Now we will study the expected value and the variance of a standard normal variable.

Corollary

Corollary 32. *If $Z \sim N(0,1)$, then $\mathbb{E}Z = 0, \mathbf{Var}(Z) = 1$.*

Proof. Left as an exercise for the reader. ■

To find the distribution function of the standard normal distribution, we need to integrate the PDF function. In particular we define

$$F_Z(x) = \Phi(x) = \mathbb{P}(Z \leq x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left\{-\frac{u^2}{2}\right\}.$$

And based on that, here are some properties of the Φ function:

(i) $\lim_{x\to+\infty} \Phi(x) = 1, \lim_{x\to-\infty} \Phi(x) = 0$;

(ii) $\Phi(0) = \frac{1}{2}$;

(iii) $\Phi(-x) = 1 - \Phi(x)$, for all $x \in \mathbb{R}$.

Indeed, $\Phi(x)$ is a distribution function.

In general, we can obtain any normal random variable variable by shifting and scaling a standard normal random variable. In particular, define

$$X = \sigma Z + \mu, \quad \text{where } \sigma > 0$$

Then

$$\mathbb{E}X = \sigma \mathbb{E}Z + \mu = \mu,$$
$$\mathbf{Var}(X) = \sigma^2 \mathbf{Var}(Z) = \sigma^2.$$

And we say that X is a normal random variable with mean μ and variance σ^2. We write $X \sim N(\mu, \sigma^2)$.

Definition

Definition 35. *If Z is a standard normal random variable and $X = \sigma Z + \mu$, then X is a normal random variable with mean μ and variance σ^2, i.e*

$$X \sim N(\mu, \sigma^2).$$

Conversely, if $X \sim N(\mu, \sigma^2)$, then the random variable defined by $Z = \dfrac{X-\mu}{\sigma}$ is a standard random variable, so

$$\begin{aligned} F_X(x) &= \mathbb{P}(X \leq x) \\ &= \mathbb{P}(\sigma Z + \mu \leq x), \quad \text{where } Z \sim N(0,1) \\ &= \mathbb{P}\left(Z \leq \frac{x-\mu}{\sigma}\right) \\ &= \Phi\left(\frac{x-\mu}{\sigma}\right). \end{aligned}$$

And we may differentiate $F_X(x)$ to find the PDF of X:

$$f_X(x) = \frac{d}{dx} F_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}.$$

Corollary

Corollary 33. *If X is a normal random variable with mean μ and variance σ^2, $X \sim N(\mu, \sigma^2)$, then*

$$\mathbb{P}(a \leq X \leq b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right).$$

Below is an image showing the PDF of a normal variable with different parameters:

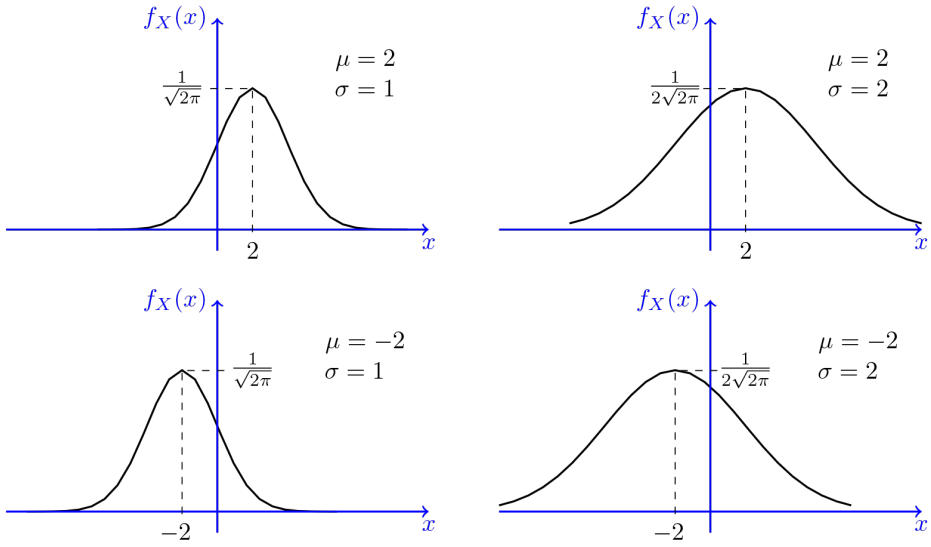


Figure 3.3: PDF of normal distribution with different parameters

An important property of the normal distribution is that a linear transformation of a normal random variable is itself a normal random variable.

Theorem

Theorem 32. *If $X \sim N(\mu_X, \sigma_X^2)$ and $Y = aX + b$ where $a, b \in \mathbb{R}$, then*

$$Y \sim N(a\mu_X + b, a^2\sigma_X^2).$$

Proof. This follows trivially from the arguments shown above. ■

3.5.4 Gamma Distribution

The gamma distribution is another widely used distribution. Its importance is largely due to its relation to exponential and normal distributions. Before introducing the gamma random variable, we need to introduce the gamma function.

Definition

Definition 36. *The gamma function, shown by $\Gamma(x)$ is an extension of the factorial function to real and complex numbers. If $n \in \{1,2,3,\cdots\}$, then*

$$\Gamma(n) = (n-1)!$$

More generally, for any positive real number α, define

$$\Gamma(\alpha) = \int_0^{+\infty} x^{\alpha-1} e^{-x} dx, \quad \text{for } \alpha > 0.$$

Below is an image of the gamma function for some positive α:

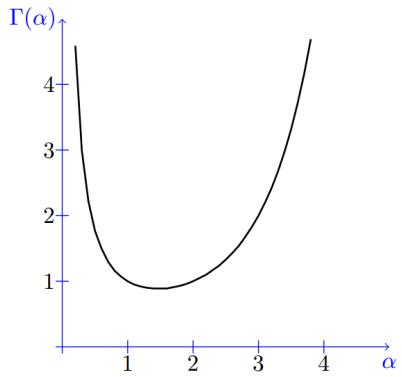


Figure 3.4: The gamma function for some real values of α

Note that when $\alpha = 1$, we have

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

Using the change of variable $x = \lambda y$, the following equation is also very useful:

$$\Gamma(\alpha) = \lambda^{\alpha} \int_0^{+\infty} y^{\alpha-1} e^{-\lambda y} dy, \quad \text{for } \alpha, \lambda > 0.$$

More properties of a gamma function follows:

(i) $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$;

(ii) $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$.

Now we may define the gamma distribution:

Definition

Definition 37. *A continuous random variable X is said to have a gamma distribution with parameters $\alpha > 0$ and $\lambda > 0$, shown as $X \sim Gamma(\alpha, \lambda)$, if its PDF is given by*

$$f_X(x) = \begin{cases} \dfrac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)} & x > 0 \\ 0 & otherwise \end{cases}$$

If $\alpha = 1$, we get $Gamma(1, \lambda) = Exponential(\lambda)$, which is a special case. More generally, if we take the sum of n independent $Exponential(\lambda)$ random variables, we will get a $Gamma(n, \lambda)$ random variable. It can be proved by using moment generating functions. The figure below shows the PDF of a gamma distribution with different parameters α:

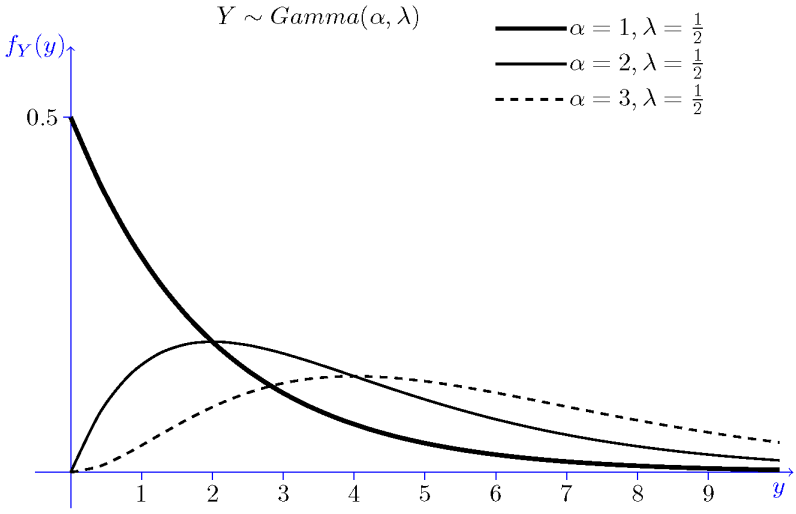


Figure 3.5: Gamma Distribution with Different Parameters

Corollary

Corollary 34. *If $X \sim Gamma(\alpha, \lambda)$, then*

$$\mathbb{E}X = \frac{\alpha}{\lambda}, \mathbf{Var}(X) = \frac{\alpha}{\lambda^2}.$$

3.5.5 Weibull Distribution

The Weibull distribution is widely used in engineering practice due to its versatility. It was originally proposed for the interpretation of fatigue data, but now its use has been extended to many other engineering problems. In particular, it is widely used in the field of life phenomena as the distribution of the lifetime of some object, especially when the "weakest link" model is appropriate for the object. That is, consider

an object consisting of many parts, and suppose that the object experiences death (failure) when any of its parts fail. It has been shown (both theoretically and empirically) that under these conditions a Weibull distribution provides a close approximation to the distribution of the lifetime of the item.

Definition

Definition 38. *A continuous random variable is said to have a Weibull distribution if its distribution function is given by*

$$F(x) = \begin{cases} 0 & x \leq \nu \\ 1 - \exp\left\{ -\left(\dfrac{x-\nu}{\alpha} \right)^{\beta} \right\} & x > \nu \end{cases}$$

By differentiating the function above, we get the PDF of a Weibull random variable with parameters ν, α, β:

$$f(x) = \begin{cases} 0 & x \leq \nu \\ \dfrac{\beta}{\alpha} \left(\dfrac{x-\nu}{\alpha} \right)^{\beta-1} \exp\left\{ -\left(\dfrac{x-\nu}{\alpha} \right)^{\beta} \right\} & x > \nu \end{cases}$$

3.5.6 Cauchy Distribution

Definition

Definition 39. *A continuous random variable is said to have a Cauchy Distribution with parameter $\theta \in \mathbb{R}$, if its PDF is given by*

$$f(x) = \frac{1}{\pi} \frac{1}{1 + (x-\theta)^2} \quad x \in \mathbb{R}$$

Below is an example of Cauchy distribution.

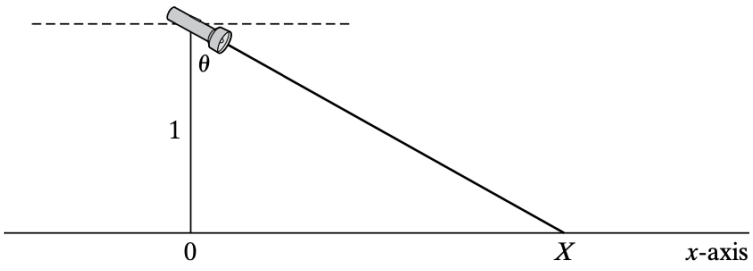


Figure 3.6

Suppose that a narrow beam flashlight is spun around its center, which is located a unit distance from the x-axis. Consider the point X at which the beam interests the X axis when the flashlight has stopped

spinning, as shown in figure 3.6.

The point X is determined by the angle θ between the flashlight and the y-axis, which from the physical situation, appears to be uniformly distributed between $-\pi/2$ and $\pi/2$/ The distribution function of X is thus given by

$$\begin{aligned} F(x) &= \mathbb{P}\{X \leq x\} \\ &= \mathbb{P}\{\tan\theta \leq x\} \\ &= \mathbb{P}\{\theta \leq \tan^{-1} x\} \\ &= \frac{1}{2} + \frac{1}{\pi}\tan^{-1} x. \end{aligned}$$

Where we know that

$$\mathbb{P}\{\theta \leq a\} = \frac{a-(-\pi/2)}{\pi} = \frac{1}{2} + \frac{a}{\pi} \quad -\frac{\pi}{2} < a < \frac{\pi}{2}$$

Hence the density function of X is given by

$$f(x) = \frac{d}{dx}F(x) = \frac{1}{\pi(1+x^2)} \quad x \in \mathbb{R}$$

and we see that X has the Cauchy distribution.

Chapter 4

Multiple Random Variables

Multiple Random Variables

Definition

Definition 40. *The vector* $\mathbf{X} = (X_1, X_2, \cdots, X_n)$ *defined on* $(\Omega, \mathscr{F}, \mathbb{P})$ *into* $\mathbb{R}^n$ *by*

$$\mathbf{X}(\omega) = \{X_1(\omega), X_2(\omega), \cdots, X_n(\omega)\}, \omega \in \Omega$$

is called an n-dimensional random variable if the inverse image of every n-dimensional interval is also in $\mathscr{F}$.

Let $\mathbf{X} = (X_1, X_2, \cdots, X_n)$ be a vector of random variables, and $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$, we define

$$F_{\mathbf{X}}(\boldsymbol{x}) = \mathbb{P}\{X_1 \leq x_1; X_2 \leq x_2; \cdots; X_n \leq x_n\}$$

And we claim the following:

① $F_{\mathbf{X}}$ is non-decreasing in every argument $x_1, x_2, \cdots, x_n$;

② $F_{\mathbf{X}}$ is right continuous in all arguments $x_1, x_2, \cdots, x_n$;

③ $F_{\mathbf{X}}(-\infty, x_2, \cdots, x_n) = F_{\mathbf{X}}(x_1, -\infty, x_3, \cdots, x_n) = \cdots = 0$;

④ $F_{\mathbf{X}}(+\infty, +\infty, \cdots, +\infty) = 1$;

⑤ Let $n = 2$, $\forall \varepsilon_1, \varepsilon_2 > 0$, we have

$$F_{\mathbf{X}}(x_1 + \varepsilon, x_2 + \varepsilon) - F_{\mathbf{X}}(x_1 + \varepsilon, x_2) - F_{\mathbf{X}}(x_1, x_2 + \varepsilon_2) + F_{\mathbf{X}}(x_1, x_2) \geq 0$$

This idea can also be generalized for $n \geq 2$.

For better notation and simplification, we will discuss the case when $n = 2$ from now on.

Definition

Definition 41. *A 2-dimensional random variable (X,Y) is said to be of the discrete type if it takes on pairs of values belonging to a countable set A with probability 1. We define $p_{ij} = (x_i, y_j)$ for every pair, and define*

$$p_{ij} = \mathbb{P}\{X = x_i; Y = y_j\}$$

to be the joint probability mass function of (X,Y), the distribution function is given by

$$F(x,y) = \sum_{(i,j)\in B} p_{ij}, \quad B = \{(i,j) : x_i \leq x; y_j \leq y\}$$

We also give the definition for the continuous type:

Definition

Definition 42. *A 2-dimensional random variable (X,Y) is said to be of the continuous type if there exists a non-negative function f, such that for every pair $(x,y) \in \mathbb{R}^2$, we have*

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

where $F(x,y)$ is the distribution function of (X,Y) and f is the joint probability density function of (X,Y).

We know that $f(x,y) \geq 0$, also

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)dudv = 1$$

And in particular if f is twice continuous at (x,y), then we have

$$\frac{\partial^2 F(x,y)}{\partial x \partial y} = f(x,y)$$

Theorem

Theorem 33. *If f is a non-negative function satisfying $\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)dxdy = 1$, then f is the joint probability density function for some random variable.*

Suppose a 2-dimensional random variable (X,Y), assume it is of the continuous type, then we define

$$F_1(x) = \mathbb{P}(X_1 \leq x) = F(x,+\infty) = \lim_{y\to\infty} F(x,y)$$

and

$$\lim_{y\to\infty} F(x,y) = \lim_{y\to\infty} \mathbb{P}\{X \leq x, Y \leq y\}.$$

Likewise,

$$F_2(y) = F(+\infty, y) = \lim_{x\to\infty} F(x,y)$$

and

$$\lim_{x\to\infty} F(x,y) = \lim_{x\to\infty} \mathbb{P}\{X \le x, Y \le y\}$$

Also we have

$$f_1(x) = \int_{-\infty}^{+\infty} f(x,y)dy; f_2(y) = \int_{-\infty}^{+\infty} f(x,y)dx$$

We have $f_1(x), f_2(y) \ge 0$, and $f_1(x)$ is called the marginal probability density function of X, and $f_2(y)$ is called the marginal probability density function of Y.

Now for discrete type, assume $X = \{x_1, x_2, \cdots\}, Y = \{y_1, y_2, \cdots\}$, then

$$F(x, +\infty) = \sum \sum_{(i,j)\in B} p_{ij}, \;\; B = \{(i,j) : X_i \le x, y_j \le +\infty\}$$

$$f_1(x_i) = \mathbb{P}(X = x_i) = \sum_{j=1}^{+\infty} \mathbb{P}(X = x_i), Y = y_j) = \sum_{j=1}^{+\infty} p_{ij} = p_{i\cdot}$$

Likewise,

$$f_2(y_j) = \mathbb{P}(Y = y_j) = \sum_{i=1}^{+\infty} p_{ij} = p_{\cdot j}$$

Likewise, $p_{i\cdot}$ is called the marginal probability mass function of X, and $p_{\cdot j}$ is called the marginal probability mass function of Y.

Example: Consider a circle of radius R centered at the origin. A point is randomly chosen on the circle, and we say the the point is uniformly distributed within the circle. That is, assume the point has coordinates (x,y), then the joint distribution of the random variable X, Y is given by

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{\pi R^2} : x^2 + y^2 \le R^2 \\ 0 : \text{otherwise} \end{cases}$$

Compute the marginal density functions of X, Y; compute the probability that D, the distance from the origin of the point selected, is less than or equal to a, where a is a given constant. Finally find $\mathbb{E}(D)$.

solution: The marginal distribution of X is given by

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{+\infty} \frac{1}{\pi R^2} dy \\ &= \frac{1}{\pi R^2} \int_{-\sqrt{R^2-x^2}}^{\sqrt{R^2-x^2}} dy \\ &= \frac{2}{\pi R^2} \sqrt{R^2 - x^2}, \text{where } x^2 \le R^2. \end{aligned}$$

and is equal to 0 otherwise. Likewise, we have

$$f_Y(y) = \frac{2}{\pi R^2} \sqrt{R^2 - y^2}, \text{where } y^2 \le R^2.$$

and is equal to 0 otherwise.

Since the distace is always non-negative, thus we have

$$\begin{aligned}\mathbb{P}\{\sqrt{X^2+Y^2} \le a\} &= \mathbb{P}\{X^2+Y^2 \le a^2\} \\ &= \iint_{x^2+y^2\le a^2} f(x,y)dxdy \\ &= \frac{1}{\pi R^2}\iint_{x^2+y^2\le a^2} dxdy \\ &= \frac{\pi a^2}{\pi R^2} \\ &= \left(\frac{a}{R}\right)^2.\end{aligned}$$

So we know that $F_D(a) := \frac{a^2}{R^2}$, so

$$f_D(a) = \frac{2a}{R^2}, 0 \le a \le R$$

Hence

$$\mathbb{E}D = \frac{2}{R^2}\int_0^R a^2 da = \frac{2R}{3}$$

Example: Suppose (X,Y) be jointly distributed with joint probability density function given by $f(x,y) = 2, 0 < x < y < 1$, and $f(x,y) = 0$ otherwise. Then we have

$$f_1(x) = \int_x^1 2dy = \begin{cases} 2-2x, 0 < x < 1 \\ 0, \text{otherwise} \end{cases}$$

$$f_2(y) = \int_0^y 2dx = \begin{cases} 2y, 0 < y < 1 \\ 0, \text{otherwise} \end{cases}$$

Definition

Definition 43. *Let (X,Y) be a random variable of the discrete type, if $\mathbb{P}(Y = y_j) > 0$, the function*

$$p_{i|j} = \mathbb{P}(X = x_i | Y = y_j) = \frac{\mathbb{P}(X = x_i, Y = y_j)}{\mathbb{P}(Y = y_j)}$$

for a fixed j is known as the conditional probability mass function of X given $Y = y_j$, a similar definition can also be given on Y for a fixed x_i

By definition, we also have

$$p_{i|j} = \frac{p_{ij}}{p_{\cdot j}}$$

$$F_{X|Y}(x,y) = \mathbb{P}(X \le x | Y = y) = \frac{\mathbb{P}(X \le x, Y = y)}{\mathbb{P}(Y = y)}$$

For a continuous random variable, we have

$$\begin{aligned} F_{X|Y}(x|y) &= \mathbb{P}(X \leq x | Y = y) \\ &= \lim_{\varepsilon \to 0+} \mathbb{P}(X \leq x | Y \in (y-\varepsilon, y+\varepsilon)) \\ &= \lim_{\varepsilon \to 0+} \frac{\mathbb{P}(X \leq x, Y \in (y-\varepsilon, y+\varepsilon))}{\mathbb{P}(Y \in (y-\varepsilon, y+\varepsilon))} \\ &= \lim_{\varepsilon \to 0+} \left(\frac{\int_{-\infty}^{x} \int_{y-\varepsilon}^{y+\varepsilon} f(u,v) du dv}{\int_{y-\varepsilon}^{y+\varepsilon} f_2(v) dv} \right) \\ &= \frac{\int_{-\infty}^{x} f(u,y) du}{f_2(y)} = \int_{-\infty}^{x} \frac{f(u,y)}{f_2(y)} du \end{aligned}$$

and thus

$$f_{X|Y}(x,y) = \frac{f(x,y)}{f_2(y)}$$

is called the conditional probability density function of the random variable of the continuous type.

Definition

Definition 44. *The conditional probability distribution function for a random variable of the continuous type X given $Y = y$ is given by*

$$F_{X|Y}(x|y) = \lim_{\varepsilon \to 0^+} \mathbb{P}\{X \leq x | Y \in (y-\varepsilon, y+\varepsilon]\}$$

Theorem

Theorem 34. *Let f be the joint probability density function of a random variable (X,Y) of the continuous type. Let $f_2(y)$ be the marginal probability density function of Y. At every point (x,y) which f is continuous and $f_2(Y) > 0$ and also continuous, the conditional probability density function of X given $Y = y$ is given by*

$$f_{X|Y}(x|y) = \frac{f(x,y)}{f_2(y)}$$

and the same idea applies for $f_{Y|X}(y|x)$.

Note that

$$\int_{-\infty}^{x} f(u,y) du = f_2(y) F_{X|Y}(x|y)$$

so we have

$$F_1(x) = \int_{-\infty}^{+\infty} \left[\int_{-\infty}^{x} f(u,y) du \right] dy = \int_{-\infty}^{+\infty} f_2(y) F_{X|Y}(x|y) dy$$

where $F_1(x)$ is the marginal distribution function of X.

Example: Consider random variable (X,Y) which has a joint distribution function given by

$$f(x,y) = \begin{cases} 2, 0 < x < y < 1 \\ 0, \text{otherwise} \end{cases}$$

Then we can find

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_1(x)} = \frac{1}{1-x}, \;\; x < y < 1$$

also

$$f_{X|Y}(x|y) = \frac{f(x,y)}{f_2(x)} = \frac{1}{y}, \;\; 0 < x < y$$

Now we also need to care about the bounds, since they might be tricky:

$$\mathbb{P}\left\{ Y \geq \frac{1}{2} \middle| x = \frac{1}{2} \right\} = \int_{\frac{1}{2}}^{1} \frac{1}{1-\frac{1}{2}} dy = 1$$

$$\mathbb{P}\left\{ X \geq \frac{1}{3} \middle| y = \frac{2}{3} \right\} = \int_{\frac{1}{3}}^{\frac{2}{3}} \frac{1}{\frac{2}{3}} dx = \frac{1}{2}.$$

Independent Random Variables

Definition

Definition 45. *Given $X = \{x_1, x_2, \cdots\}, Y = \{x_1, x_2, \cdots\}$, we say that X, Y are independent, if*

$$\mathbb{P}(X = x_i, Y = y_j) = \mathbb{P}(X = x_i)\mathbb{P}(Y = y_j)$$

for all $x_i \in X, y_j \in Y$.

We can also say that X, Y are independent if and only if

$$F(x,y) = F_1(x)F_2(y) \text{ for all } (x,y) \in \mathbb{R}^2$$

Corollary

Corollary 35. *Let X, Y be independent random variables, then we have $F_{Y|X}(y|x) = F_Y(y)$ for all y and $F_{X|Y}(x|y) = F_X(x)$ for all x.*

Example 1: A man and a woman decide to meet at a certain location. If each of them independently arrives at a time uniformly distributed between 12 noon and 1 P.M., find the probability that the first to arrive has to wait longer than 10 minutes.

Solution: If we let X, Y denote, respectively the time past 12 that the man an the woman arrive, then X, Y are independent random variables and each of which is uniformly distributed over $[0, 60]$. The desired probability is given by

$$\mathbb{P}\{X + 10 < Y\} + \mathbb{P}\{Y + 10 < X\}$$

And by symmetry, is just $2\mathbb{P}\{X + 10 < Y\}$, thus we have

$$\begin{aligned} 2\mathbb{P}\{X + 10 < Y\} &= 2 \iint_{x+10<y} f(x,y)dxdy \\ &= 2 \iint_{x+10<y} f_X(x) f_Y(y)dxdy \\ &= 2 \int_{10}^{60} \int_{0}^{y-10} \left(\frac{1}{60}\right)^2 dxdy \\ &= \frac{25}{36}. \end{aligned}$$

Example 2: (Buffon's Needle Problem) A table is ruled with equidistant parallel lines with a distance D apart. A needle of length L where $L \leq D$ is randomly thrown on the table. What is the probability that the needle will intersect one of the lines (the other probability will be that the needle will be completely contained in the strip between two lines)?

Solution:

Let X be the distance from the middle point of the needle to the nearest parallel line, and θ be the angle of the needle and the projected line of length X. Then by the construction, the needle will intersect the line if and only if

$$\frac{X}{\cos(\theta)} < \frac{L}{2}$$

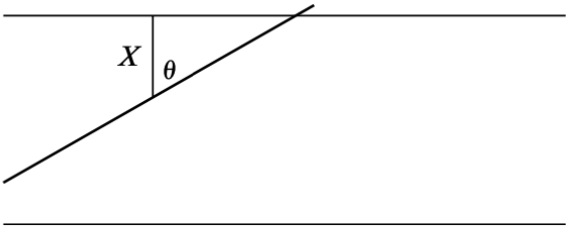


FIGURE 6.2

Figure 4.1: The graph illustrating when the needle will intersect

And we see that $X \in [0, D/2]$ and $\theta \in [0, \pi/2]$, so we may assume that they are independent, and uniformly distributed throughout their ranges, hence we have

$$\begin{aligned}
\mathbb{P}\left\{X < \frac{L}{2}\cos(\theta)\right\} &= \iint_{x<\frac{L}{2}\cos(y)} f_X(x) f_\theta y dx dy \\
&= \frac{4}{\pi D}\int_0^{\pi/2}\int_0^{\frac{L}{2}\cos(y)} dxdy \\
&= \frac{2L}{\pi D}.
\end{aligned}$$

Example 3: If random variables X, Y, Z are independent and uniformly distributed over $[0,1]$, compute $\mathbb{P}\{X \geq YZ\}$.

Solution: Since

$$f_{X,Y,Z}(x,y,z) = f_X(x) f_Y(y) f_Z(z) = 1,\ (x,y,z) \in [0,1]^3,$$

so we have

$$\begin{aligned}
\mathbb{P}\{X \geq YZ\} &= \iiint_{x\geq yz} f_{X,Y,Z}(x,y,z) dxdydz \\
&= \int_0^1\int_0^1\int_{yz}^1 dxdydz \\
&= \frac{3}{4}.
\end{aligned}$$

Suppose X, Y are independent random variables, we also want to find the probability distribution of $X+Y$.

We have

$$\begin{aligned} F_{X+Y}(a) &= \mathbb{P}\{X+Y \le a\} \\ &= \iint_{x+y\le a} f_X(x) f_Y(y) dxdy \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{a-y} f_X(x) f_Y(y) dxdy \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{a-y} f_X(x) dx f_Y(y) dy \\ &= \int_{-\infty}^{+\infty} F_X(a-y) f_Y(y) dy \end{aligned}$$

The cumulative distribution function F_{X+Y} is called the convolution of the distributions F_X and F_Y, by differentiating the equation, we get

$$f_{X+Y}(a) = \int_{-\infty}^{+\infty} f_X(a-y) f_Y(y) dy$$

Example 4: Suppose X, Y are independent random variables both uniformly distributed on $[0,1]$, find the probability density function of $X+Y$.

Solution: We know that

$$f_X(a) = f_Y(a) = \begin{cases} 1, 0 < a < 1 \\ 0, \text{otherwise} \end{cases},$$

so we have

$$f_{X+Y}(a) = \int_0^1 f_X(a-y) f_Y(y) dy = \int_0^1 f_X(a-y) dy$$

For $0 \le a \le 1$, we have

$$f_{X+Y}(a) = \int_0^a dy = a$$

and for $1 \le a \le 2$, we have

$$f_{X+Y}(a) = \int_{a-1}^1 dy = 2-a$$

Hence

$$f_{X+Y}(a) = \begin{cases} a, 0 \le a \le 1 \\ 2-a, 1 < a < 2 \\ 0, \text{otherwise} \end{cases}.$$

Corollary

Corollary 36. *Suppose $X_1, X_2, \cdots, X_n$ are independent random variables uniformly distributed on $[0,1]$, and let $F_n(x) = \mathbb{P}\{X_1 + X_2 + \cdots + X_n \le x\}$, then*

$$F_n(x) = \frac{x^n}{n!}, \text{ when } 0 \le x \le 1$$

Proof. This can be done by using induction. ■

Based on example 4, what is the expected number of n such that $X_1+X_2+\cdots+X_n>1$? That is, we want to find

$$N := \min\{n : X_1+X_2+\cdots+X_n>1\}$$

Note that $N>n$ if and only if $X_1+X_2+\cdots+X_n\leq 1$, so

$$\mathbb{P}\{N>n\} = F_n(1) = \frac{1}{n!},\ \ n>0,$$

thus

$$\mathbb{P}\{N=n\} = \mathbb{P}\{N>n-1\} - \mathbb{P}\{N>n\} = \frac{1}{(n-1)!} - \frac{1}{n!} = \frac{n-1}{n!}$$

Therefore

$$\mathbb{E}(N) = \sum_{n=1}^{+\infty} \frac{n(n-1)}{n!} = \sum_{n=2}^{+\infty} \frac{1}{(n-2)!} = e.$$

The sum of the independent random variables can also be calculated using moment generating functions. Suppose $X_1, X_2, \cdots, X_n$ are n independent random variables and

$$Y = X_1+X_2+\cdots+X_n$$

Then

$$\begin{aligned} M_Y(s) &= \mathbb{E}[e^{sY}] \\ &= \mathbb{E}[e^{s(X_1+\cdots+X_n)}] \\ &= \mathbb{E}[e^{sX_1}e^{sX_2}\cdots e^{sX_n} \\ &= \mathbb{E}[e^{sX_1}]\cdots\mathbb{E}[e^{sX_n}] \\ &= M_{X_1}(s)\cdots M_{X_n}(s). \end{aligned}$$

This important application can be used to find the moment generating function for a binomial random variable. Suppose $X\sim Binomial(n,p)$, then we know that

$$X = X_1+X_2+\cdots+X_n$$

where $X_i\sim Bernoulli(p)$, thus

$$M_X(s) = M_{X_1}(s)\cdots M_{X_n}(s)$$

where

$$M_{X_i}(s) = \mathbb{E}[e^{sX_i}] = pe^s+1-p$$

So

$$M_X(s) = (pe^s+1-p)^n.$$

Conditional Distributions

4.3.1 Conditioning by One Variable

Recall the conditional probability:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}, \text{when } \mathbb{P}(B) > 0.$$

As another example, if we have two random variables X, Y we may write

$$\mathbb{P}(X \in C|Y \in D) = \frac{\mathbb{P}(X \in C, Y \in D)}{\mathbb{P}(Y \in D)}, \text{where } C, D \in \mathbb{R}.$$

Now suppose that we have a continuous random variable X, and we know that the event $X \in I = [a,b]$ has occurred. Call this event A, the conditional CDF of X given A, denoted by $F_{X|A}(x)$ or $F_{X|a\leq X\leq b}(x)$ is given by

$$\begin{aligned} F_{X|A}(x) &= \mathbb{P}(X \leq |A) \\ &= \mathbb{P}(X \leq x|a \leq X \leq b) \\ &= \frac{\mathbb{P}(X \leq x, a \leq X \leq b)}{P(A)} \end{aligned}$$

Now if $x < a$, then $F_{X|A}(x) = 0$, on the other hand, if $a \leq x \leq b$, we have

$$\begin{aligned} F_{X|A}(x) &= \frac{\mathbb{P}(X \leq x, a \leq X \leq b)}{\mathbb{P}(A)} \\ &= \frac{\mathbb{P}(a \leq X \leq x)}{\mathbb{P}(A)} \\ &= \frac{F_X(x) - F_X(a)}{F_x(b) - F_X(a)} \end{aligned}$$

And finally if $x > b$ then $F_{X|A}(x) = 1$, thus we obtain

$$F_{X|A}(x) = \begin{cases} 1 : x > b \\ \\ \frac{F_X(x) - F_X(a)}{F_x(b) - F_X(a)} : a \leq x \leq b \\ \\ 0 : \text{otherwise} \end{cases}$$

We assume that X is a continuous random variable, we do not need to be careful about the end points, to obtain the conditional PDF of X, we may simply differentiate $F_{X|A}(x)$, which gives

$$f_{X|A}(x) = \begin{cases} \frac{f_X(x)}{\mathbb{P}(A)} : a \leq x < b \\ \\ 0 : \text{otherwise} \end{cases}$$

The conditional expectation and variance are defined by replacing the PDF by conditional PDF in the definitions of expectation and variance. In general for a random variable x and an event A, we have the followings:

(i) $\mathbb{E}[X|A] = \int_{-\infty}^{+\infty} x f_{X|A}(x)dx$;

(ii) $\mathbb{E}[g(x)|A] = \int_{-\infty}^{+\infty} g(x) f_{X|A}(x)dx$;

(iii) $\mathbf{Var}(X|A) = \mathbb{E}[X^2|A] - (\mathbb{E}[X|A])^2$.

Example : Let $X \sim Exponential(1)$, i.e

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} : x \geq 0 \\ 0 : \text{otherwise} \end{cases}$$

where $\lambda = 1$ in this case. Find : (i) The conditional PDF and CDF of X given $X > 1$; (ii) $\mathbb{E}[X|X > 1]$; (iii) $\mathbf{Var}(X|X > 1)$.

Solution:

(i) : Let A be the event that $X > 1$, then

$$\mathbb{P}(A) = \int_1^{+\infty} e^{-x} dx = \frac{1}{e}.$$

thus

$$f_{X|X>1}(x) = \begin{cases} e^{-x+1} : x > 1 \\ 0 : \text{otherwise} \end{cases}$$

and for $x > 1$, we have

$$F_{X|A}(A) = \frac{F_X(x) - F_X(1)}{\mathbb{P}(A)} = 1 - e^{-x+1}$$

and $F_{X|A}(x) = 0$ otherwise.

(ii) : We have

$$\begin{aligned} \mathbb{E}[X|X > 1] &= \int_1^{+\infty} x f_{X|X>1}(x)dx \\ &= \int_1^{+\infty} x e^{-x+1} dx \\ &= 2. \end{aligned}$$

(iii) : We have

$$\begin{aligned} \mathbb{E}[X^2|X > 1] &= \int_1^{+\infty} x^2 f_{X|X>1}(x)dx \\ &= \int_1^{+\infty} x^2 e^{-x+1} dx \\ &= 5. \end{aligned}$$

Thus

$$\mathbf{Var}(X|X>1) = \mathbb{E}[X^2|X>1] - (\mathbb{E}[X|X>1])^2 = 1.$$

4.3.2 Discrete Conditional Distribution

Recall that for any two events E,F, the conditional probability of E given F is defined, provided that $\mathbb{P}(F) > 0$ is that

$$\mathbb{P}(E|F) = \frac{\mathbb{P}(EF)}{\mathbb{P}(F)}$$

Hence, if X,Y are discrete random variables, it's natural to define the conditional probability mass function of X given that $Y = y$ by

$$\begin{aligned} p_{X|Y}(x|y) &= \mathbb{P}\{X = x|Y = y\} \\ &= \frac{\mathbb{P}\{X = x, Y = y\}}{\mathbb{P}\{Y = y\}} \\ &= \frac{p(x,y)}{p_Y(y)} \end{aligned}$$

Similarly, the conditional probability distribution function given that $Y = y$ is defined, for all y such that $p_Y(y) > 0$ by

$$F_{X|Y}(x|y) = \mathbb{P}\{X \leq x|Y = y\} = \sum_{a \leq x} p_{X|Y}(a|y)$$

In other words, the definitions are exactly the same as in the unconditional case, except that everything is not conditional on the event that $Y = y$. If X is independent of Y, then the conditional mass function and the distribution function are the same as the respective unconditional ones. This follows because if X is independent of Y, then we have

$$\begin{aligned} p_{X|Y}(x|y) &= \mathbb{P}\{X = x|Y = y\} \\ &= \frac{\mathbb{P}\{X = x, Y = y\}}{\mathbb{P}\{Y = y\}} \\ &= \frac{\mathbb{P}\{X = x\}\mathbb{P}\{Y = y\}}{\mathbb{P}\{Y = y\}} \\ &= \mathbb{P}\{X = x\}. \end{aligned}$$

Example: Suppose that $p(x,y)$, the joint probability mass function of X and Y, is given by

$$p(0,0) = 0.4, p(0,1) = 2, p(1,0) = 0.1, p(1,1) = 0.3$$

Then find the conditional probability mass function of X given that $Y = 1$.

Solution : We first note that

$$p_Y(1) = \sum_x p(x,1) = p(0,1) + p(1,1) = 0.5$$

Hence

$$p_{x|Y}(x|y) = \frac{p(1,1)}{p_Y(1)} = \frac{2}{5}$$

and

$$p_{X|Y}(1|1) = \frac{p(1,1)}{p_Y(1)} = \frac{3}{5}.$$

Example: If X, Y are independent Poisson random variables with respective parameters λ_1 and λ_2, calculate the conditional distribution of X given that $X+Y=n$.

Solution : We calculate the conditional probability mass function of X given that $X+Y=n$ as follows:

$$\begin{aligned}\mathbb{P}\{X=k|X+Y=n\} &= \frac{\mathbb{P}\{X=k, X+Y=n\}}{\mathbb{P}\{X+Y+n\}} \\ &= \frac{\mathbb{P}\{X=k, Y=n-k\}}{\mathbb{P}\{X+Y=n\}} \\ &= \frac{\mathbb{P}\{X=k\}\mathbb{P}\{Y=n-k\}}{\mathbb{P}\{X+Y=n\}}.\end{aligned}$$

Recall that $X+Y$ has a Poisson distribution with parameter $\lambda_1+\lambda_2$, we see that the preceding equals

$$\begin{aligned}\mathbb{P}\{X=k|X+Y=n\} &= \frac{e^{-\lambda_1}\lambda_1^k}{k!}\frac{e^{-\lambda_2}\lambda_2^{n-2}}{(n-k)!}\left[\frac{e^{-(\lambda_1+\lambda_2)}(\lambda_1+\lambda_2)^n}{n!}\right]^{-1} \\ &= \frac{n!}{(n-k)!k!}\frac{\lambda_1^k\lambda_2^{n-k}}{(\lambda_1+\lambda_2)^n} \\ &= \binom{n}{k}\left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^k\left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)^{n-k}.\end{aligned}$$

In other words, the conditional distribution of X given that $X+Y=n$ is the binomial distribution with parameters n and $\lambda_1/(\lambda_1+\lambda_2)$.

Example : Consider n independent trails, with each trail being a success with a probability p. Given a total of k success, show that all possible orderings of the k successes and $n-k$ failures are equally likely.

Solution : We want to show that given a total of k success, each of the $\binom{n}{k}$ possible orderings of k success and $n-k$ failures is equally likely. Let X denote the number of successes, and consider any ordering of k success and $n-k$ failures, say $\boldsymbol{o} = (s,s,f,f,\cdots,f)$, then

$$\begin{aligned}\mathbb{P}\{\boldsymbol{o}|X=K) &= \frac{\mathbb{P}\{\boldsymbol{o}, X=k\}}{\mathbb{P}\{X=k\}} \\ &= \frac{\mathbb{P}(\boldsymbol{o})}{\mathbb{P}\{X=k\}} \\ &= \frac{p^k(1-p)^{n-k}}{\binom{n}{k}p^k(1-p)^{n-k}} \\ &= \frac{1}{\binom{n}{k}}.\end{aligned}$$

Example : Consider two random variables X and Y with joint probability mass function given by

	$Y=2$	$Y=4$	$Y=5$
$X=1$	$\frac{1}{12}$	$\frac{1}{24}$	$\frac{1}{24}$
$X=2$	$\frac{1}{6}$	$\frac{1}{12}$	$\frac{1}{8}$
$X=3$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{12}$

Find : (a) $\mathbb{P}\{X \leq 2, Y \leq 4\}$; (b) The marginal probability mass functions of X and Y; (c) $\mathbb{P}\{Y=2|X=1\}$; (d) Are X, Y independent?

Solution :

(a) : By definition, we have

$$\begin{aligned}\mathbb{P}\{X \leq 2, Y \leq 4\} &= \sum_{i\leq 2, j\leq 4} \mathbb{P}\{X=i, Y=j\} \\ &= p_{12}+p_{14}+p_{22}+p_{24} \\ &= \frac{1}{12}+\frac{1}{24}+\frac{1}{6}+\frac{1}{12} \\ &= \frac{3}{8}.\end{aligned}$$

(b) : The marginal pmf of X is given by:

$$\begin{aligned}p_{1\cdot} &= p_{12}+p_{14}+p_{15} = \frac{1}{12}+\frac{1}{24}+\frac{1}{24} = \frac{4}{24} \\ p_{2\cdot} &= p_{22}+p_{24}+p_{25} = \frac{1}{6}+\frac{1}{12}+\frac{1}{8} = \frac{9}{24} \\ p_{3\cdot} &= p_{32}+p_{34}+p_{35} = \frac{1}{4}+\frac{1}{8}+\frac{1}{12} = \frac{12}{24}\end{aligned}$$

and thus we have

$$f_X(x) = \begin{cases} \frac{4}{24} : X=1 \\ \frac{9}{24} : X=2 \\ \frac{12}{24} : X=3 \\ 0 : \text{otherwise} \end{cases}$$

Same idea may be applied on Y, The marginal pmf of X is given by:

$$\begin{aligned}p_{\cdot 2} &= p_{12}+p_{22}+p_{32} = \frac{1}{12}+\frac{1}{6}+\frac{1}{4} = \frac{12}{24} \\ p_{\cdot 4} &= p_{14}+p_{24}+p_{34} = \frac{1}{24}+\frac{1}{12}+\frac{1}{8} = \frac{6}{24} \\ p_{\cdot 5} &= p_{15}+p_{25}+p_{35} = \frac{1}{24}+\frac{1}{8}+\frac{1}{12} = \frac{6}{24}\end{aligned}$$

and thus we have

$$f_Y(y) = \begin{cases} \frac{12}{24} : Y=2 \\ \frac{6}{24} : Y=4 \\ \frac{6}{24} : Y=5 \\ 0 : \text{otherwise} \end{cases}$$

(c) : By definition, we have

$$\begin{aligned}\mathbb{P}\{Y=2|X=1\} &= \frac{\mathbb{P}\{X=1,Y=2\}}{\mathbb{P}\{X=1\}} \\ &= \frac{p_{12}}{p_{1\cdot}} \\ &= \frac{\frac{1}{12}}{\frac{1}{6}} \\ &= \frac{1}{2}.\end{aligned}$$

(d) : To check that whether X,Y are independent, we need to check

$$\mathbb{P}\{X=x,Y=y\} = \mathbb{P}\{X=x\}\mathbb{P}\{Y=y\} \qquad (*)$$

For all possible pairs of X,Y. We find that

$$\mathbb{P}\{X=2,Y=2\} = \frac{1}{6}$$

But

$$\mathbb{P}\{X=2\}\mathbb{P}\{Y=2\} = \frac{3}{8}\times\frac{1}{2} = \frac{3}{16}$$

Which means this pair does not satisfy equation $(*)$, thus they are not independent.

Example : Suppose that the number of customers visiting a fast food restaurant in a given day is $N \sim Poisson(\lambda)$. Further assume that each customer purchases a drink with probability p, independently from other customers and independently from the value N. Let X be the number of customers who purchase drinks, let Y be the number of customer that do not purchase drinks, so $X+Y=N$. Find : (a) The marginal PMFs of X and Y; (b) The joint PMF of X and Y; (c) Are X and Y independent? (d) Find $\mathbb{E}[X^2Y^2]$.

Solution:

(a) : Based on what we are given, clearly

$$X|(N=n) \sim Binomial(n,p); Y|(N=n) \sim Binomial(n,1-p)$$

Then by the law of total probability, we have

$$\begin{aligned}
\mathbb{P}_X(k) &= \sum_{n=0}^{+\infty} \mathbb{P}\{X=k|N=n\}\mathbb{P}_N(n) \\
&= \sum_{n=k}^{+\infty} \binom{n}{k} p^k (1-p)^{n-k} \exp(-\lambda)\frac{\lambda^n}{n!} \\
&= \sum_{n=k}^{+\infty} \frac{p^k(1-p)^{n-k}\exp(-\lambda)\lambda^n}{k!(n-k)!} \\
&= \frac{\exp(-\lambda)(\lambda p)^k}{k!} \sum_{n=k}^{+\infty} \frac{(\lambda(1-p))^{n-k}}{(n-k)!} \\
&= \frac{\exp(-\lambda)(\lambda p)^k}{k!} \exp(\lambda(1-p)) \\
&= \frac{\exp(-\lambda p)(\lambda p)^k}{k!}, \text{for } k = 0,1,2,\cdots
\end{aligned}$$

and thus we conclude that

$$X \sim Poisson(\lambda p); Y \sim Poisson(\lambda(1-p))$$

Example : Suppose tossing a coin with $\mathbb{P}(H) = p$, repeatedly toss the coin until there are two consecutive heads. Let X denote the total number of coin tosses, find $\mathbb{E}X$.

Solution : Suppose $\mathbb{E}X = \mu$, we first condition on the result of the first toss, then

$$\mu = \mathbb{E}X = \mathbb{E}[X|H]\mathbb{P}(H) + \mathbb{E}[X|T]\mathbb{P}(T) = \mathbb{E}[X|H]p + (1+\mu)(1-p).$$

Since geometric distribution is "memoryless", i.e $\mathbb{E}X$ should be a fixed value regardless of the number of fails it gave. So if the first toss is a tail, it is then considered as a "fail" and start over again, but the number of trails increase by 1, however $\mathbb{E}X$ is still fixed. So that's why we have $\mathbb{E}(X|H) = 1 + \mathbb{E}X$, and we get

$$p\mathbb{E}X = \mathbb{E}[X|H]p + 1 - p \qquad (*)$$

Then, in order to find $\mathbb{E}[X|H]$, we condition on the second trail:

$$\begin{aligned}
\mathbb{E}[X|H] &= \mathbb{E}[X|HH]\mathbb{P}(H) + \mathbb{E}[X|HT]\mathbb{P}(T) \\
&= \underbrace{\mathbb{E}[X|HH]}_{\text{success with 2 tosses}} \cdot p + \underbrace{\mathbb{E}[X|HT]}_{\text{a fail}}(1-p) \qquad (**) \\
&= 2p + (2+\mathbb{E}X)(1-p) = 2 + (1-p)\mathbb{E}X
\end{aligned}$$

Now combine equations $(*), (**)$, we have

$$p\mathbb{E}X = (2p + (2+\mathbb{E}X)(1-p))p + 1 - p$$

Thus

$$\mathbb{E}X = \frac{1+p}{p^2}.$$

4.3.3 Continuous Conditional Distribution

Definition

Definition 46. *If X,Y have a joint probability density function $f(x,y)$, then the conditional probability density function of X given that $Y = y$ is defined, for all values of y such that $f_Y(y) > 0$, by*

$$f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)}$$

We may multiply the left-hand-side by dx and the right hand side by $(dxdy)/dy$ and we will get

$$\begin{aligned} f_{X|Y}(x|y)dx &= \frac{f(x,y)dxdy}{f_Y(y)dy} \\ &\approx \frac{\mathbb{P}\{x \le X \le x+dx, y \le Y \le y+dy\}}{\mathbb{P}\{y \le Y \le y+dy\}} \\ &= \mathbb{P}\{x \le X \le x+dx | y \le Y \le y+dy\} \end{aligned}$$

The use of conditional densities allows us to define conditional probabilities of events associated with one random variable when we are given the value of a second random variable. That is, if X and Y are jointly continuous, then for any set A,

$$\mathbb{P}\{X \in A | Y = y\} = \int_A f_{X|Y}(x|y)dx$$

In particular, by letting $A = (-\infty, a]$, we can define the conditional cumulative distribution function of X given that $Y = y$ by

$$F_{X|Y}(a|y) \equiv \mathbb{P}\{X \le a | Y = y\} = \int_{-\infty}^{a} f_{X|Y}(x|y)dx$$

For two jointly continuous random variables X,Y, we can define the following conditional concepts:

PROPERTIES OF JOINTLY CONDITIONAL DISTRIBUTION

(i) The marginal distribution of y, denoted by $f_Y(y)$ is given by

$$f_Y(y) = \int_{-\infty}^{+\infty} f_{X.Y}(x,y)dx$$

(ii) The conditional PDF of X given $Y = y$:

$$f_{X|Y}(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)}$$

(iii) The conditional probability that $X \in A$ given $Y = y$:

$$\mathbb{P}\{X \in A|Y = y\} = \int_A f_{X|Y}(x|y)dx$$

(iv) The conditional CDF of X given $Y = y$:

$$F_{X|Y}(x|y) = \mathbb{P}\{X \leq x|Y = y\} = \int_{-\infty}^{x} f_{X|Y}(x|y)dx$$

(v) The expected value of x given $Y = y$:

$$\mathbb{E}[X|Y = y] = \int_{-\infty}^{+\infty} x f_{X|Y}(x|y)dx$$

(vi) Conditional LOTUS:

$$\mathbb{E}[g(x)|Y = y] = \int_{-\infty}^{+\infty} g(x) f_{X|Y}(x|y)dx$$

(vii) Conditional variance of X given $Y = y$:

$$\mathbf{Var}(X|Y = y) = \mathbb{E}[X^2|Y = y] - (\mathbb{E}[X|Y = y])^2$$

Example: Let X, Y be jointly distributed as

$$f_{XY}(x,y) = \begin{cases} \frac{x^2}{4} + \frac{y^2}{6} + \frac{xy}{6} : 0 \leq x \leq 1, 0 \leq y \leq 2 \\ \\ 0 : \text{otherwise} \end{cases}$$

Find : (i) The conditional PDF of X given $Y = y$ where $0 \leq y \leq 2$; (ii) $\mathbb{P}\{X < 0.5|Y = y\}$; (iii) $\mathbb{E}[X|Y = 1]$; (iv) $\mathbf{Var}(X|Y = 1)$.

Solution:
(i) : We first compute $f_Y(y)$, given by

$$f_Y(y) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y)dx = \int_0^1 \frac{x^2}{4} + \frac{y^2}{4} + \frac{xy}{6}dx = \frac{1}{12} + \frac{y^2}{4} + \frac{y}{12}.$$

Then we know that

$$\begin{aligned} f_{X|Y}(x,y) &= \frac{f_{XY}(x,y)}{f_Y(y)} \\ &= \frac{\frac{x^2}{4}+\frac{y^2}{4}+\frac{xy}{6}}{\frac{1}{12}+\frac{y^2}{4}+\frac{y}{12}} \\ &= \frac{3x^2+3y^2+2xy}{3y^2+y+1} \end{aligned}$$

and thus for $0 \leq y \leq 2$ we obtain

$$f_{X|Y}(x|y) = \begin{cases} \frac{3x^2+3y^2+2xy}{3y^2+y+1} : 0 \leq x \leq 1 \\ \\ 0 : \text{otherwise} \end{cases}$$

(ii) : We have

$$\begin{aligned} \mathbb{P}\left(X < \frac{1}{2} \middle| Y = y\right) &= \int_0^{\frac{1}{2}} \frac{3x^2+3y^2+2xy}{3y^2+y+1} dx \\ &= \frac{\frac{3}{2}y^2+\frac{y}{4}+\frac{1}{8}}{3y^2+y+1}. \end{aligned}$$

(iii) : We have

$$\begin{aligned} \mathbb{E}[X|Y=1] &= \int_{\infty}^{+\infty} x f_{X|Y}(x|y=1) dx \\ &= \int_0^1 x \frac{3x^2+3y^2+2xy}{3y^2+y+1}\bigg|_{y=1} dx \\ &= \int_0^1 x \frac{3x^2+3+2x}{3+1+1} dx \\ &= \frac{1}{5}\int_0^1 (3x^3+2x^2+3x) dx \\ &= \frac{7}{12}. \end{aligned}$$

(iv) We have

$$\begin{aligned} \mathbb{E}[X^2|y=1] &= \int_{-\infty}^{+\infty} x^2 f_{X|Y}(x|y=1) dx \\ &= \frac{1}{5}\int_0^1 (3x^4+2x^3+3x^2) dx \\ &= \frac{21}{50}. \end{aligned}$$

So we have

$$\begin{aligned}\mathbf{Var}(X|Y=1) &= \mathbb{E}[X^2|Y=1] - (\mathbb{E}[X|Y=1])^2 \\ &= \frac{21}{50} - \left(\frac{7}{12}\right)^2 \\ &= \frac{287}{3600}.\end{aligned}$$

Example: The joint density function of X and Y is given by

$$f(x,y) = \begin{cases} x(2-x-y) : (x,y) \in (0,1)^2 \\ 0 : \text{otherwise} \end{cases}$$

Compute the conditional density of X given that $Y = y$, $0 < y < 1$.

Solution: For $(x,y) \in (0,1)^2$, we have

$$\begin{aligned}f_{X|Y}(x|y) &= \frac{f(x,y)}{f_Y(y)} \\ &= \frac{f(x,y)}{\int_{-\infty}^{+\infty} f(x,y)dx} \\ &= \frac{x(2-x-y)}{\int_0^1 x(2-x-y)dx} \\ &= \frac{x(2-x-y)}{\frac{2}{3}-\frac{y}{2}} \\ &= \frac{6x(2-x-y)}{4-3y}.\end{aligned}$$

Example : Suppose that the joint density of X and Y is given by

$$f(x,y) = \begin{cases} \frac{e^{-x/y}e^{-y}}{y} : (x,y) \in (0,+\infty)^2 \\ 0 : \text{otherwise} \end{cases}$$

Find $\mathbb{P}\{X > 1|Y = y\}$.

Solution : We obtain the conditional density of X given that $Y = y$:

$$\begin{aligned}f_{X|Y}(x|y) &= \frac{f(x,y)}{f_Y(y)} \\ &= \frac{e^{-x/y}e^{-y}/y}{e^{-y}\int_0^{+\infty}(1/y)e^{-x/y}dx} \\ &= \frac{1}{y}e^{-x/y}\end{aligned}$$

Hence,

$$\begin{aligned}\mathbb{P}\{X > 1 | Y = y\} &= \int_1^{+\infty} \frac{1}{y} e^{-x/y} dx \\ &= -e^{-x/y} \Big|_1^{+\infty} \\ &= e^{-1/y}.\end{aligned}$$

If X and Y are independent continuous random variables, the conditional density of X given that $Y = y$ is just the unconditional density of X.

Order Statistics

Definition

Definition 47. *Let $X_1, X_2, \cdots, X_n$ be n independent and identically distributed continuous random variables having a common density f and distribution function F, define*

$$\begin{aligned} X_{(1)} &= \text{the smallest of } X_1, X_2, \cdots, X_n \\ X_{(2)} &= \text{the second smallest of } X_1, X_2, \cdots, X_n \\ &\vdots \\ X_{(j)} &= \text{the } j\text{th smallest of } X_1, X_2, \cdots, X_n \\ &\vdots \\ X_{(n)} &= \text{the largest of } X_1, X_2, \cdots, X_n \end{aligned}$$

Then the ordered values $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ are known as the order statistics corresponding to the random variables $X_1, X_2, \cdots, X_n$.

The joint density function of the order statistics is obtained by noting that the order statistics $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ will take on the values $x_1 \leq x_2 \leq \cdots \leq x_n$ if and only if for some permutation $(i_1, i_2, \cdots, i_n) \in S_n$, we have

$$X_1 = x_{i_1}, X_2 = x_{i_2}, \cdots, X_n = x_{i_n}$$

Since for any permutation $(i_1, i_2, \cdots, i_n) \in S_n$,

$$\begin{aligned} &\mathbb{P}\left\{x_{i_1} - \frac{\varepsilon}{2} < X_1 < x_{i_1} + \frac{\varepsilon}{2}, \cdots, x_{i_n} < X_n < x_{i_n} + \frac{\varepsilon}{2}\right\} \\ &\approx \varepsilon^n f_{X_1, X_2, \cdots, X_n}(x_{i_1}, x_{i_2}, \cdots, x_{i_n}) \\ &= \varepsilon^n f(x_1) \cdots f(x_n) \end{aligned}$$

it follows that for $x_1 < x_2 < \cdots < x_n$,

$$\begin{aligned} &\mathbb{P}\left\{x_1 - \frac{\varepsilon}{2} < X_{(1)} < x_1 + \frac{\varepsilon}{2}, \cdots, x_n < X_{(n)} < x_n + \frac{\varepsilon}{2}\right\} \\ &\approx n! \varepsilon^n f(x_1) \cdots f(x_n) \end{aligned}$$

Dividing by ε^n and letting $\varepsilon \to 0$ yields

$$f_{X_{(1)}, \cdots, X_{(n)}}(x_1, x_2, \cdots, x_n) = n! f(x_1) f(x_2) \cdots f(x_n), \;\; x_1 < x_2 < \cdots < x_n$$

Example 1: Along a 1 mile long road are 3 people "distributed at random". Find the probability that no 2 people are less than a distance of d miles apart when $d \leq \frac{1}{2}$.

Solution : Assume the the positions of the 3 people are independent and uniformly distributed over the road, if X_i denote the position of the ith person, the desired probability is given by

$$\mathbb{P}\{X_{(i)} > X_{(i-1)} + d, i = 2, 3\}$$

and we have

$$f_{X_{(1)},X_{(2)},X_{(3)}}(x_1,x_2,x_3) = 3!,\ \ 0 < x_1 < x_2 < x_3 < 1$$

Thus

$$\begin{aligned}\mathbb{P}\{X_{(i)} > X_{(i-1)} + d, i = 2,3\} &= \iiint_{x_i > x_{j-1}+d} f_{X_{(1)},X_{(2)},X_{(3)}}(x_1,x_2,x_3)dx_1dx_2dx_3 \\ &= 3!\int_0^{1-2d}\int_{x_1+d}^{1-d}\int_{x_2+d}^{1} dx_3dx_2dx_1 \\ &= (1-2d)^3\end{aligned}$$

In fact, this method can be generalized, when n people are distributed at random over the unit interval, the desired probability is

$$[1-(n-1)d]^n,\ \text{ when }\ d \le \frac{1}{n-1}$$

The density function of the jth order statistic $X_{(j)}$ can be obtained either by integrating the joint density function or by direct reasoning as follows: In order for $X_{(j)}$ to equal x, it is necessary for $j-1$ of the n values $X_1,\cdots,X_n$ to be less than x and $n-j$ of them to be greater than x, and 1 of them to equal x. Now the probability density that any given set of $j-1$ of the X_i's are less than x, another given set of $n-j$ are all greater than x, and the remaining value is equal to x equals

$$[F(x)]^{j-1}[1-F(x)]^{n-j}f(x)$$

Hence, since there are

$$\binom{n}{j-1,n-j,1} = \frac{n!}{(n-j)!(j-1)!}$$

different partitions, it follows that the density function of $X_{(j)}$ is given by

$$f_{X_{(j)}}(x) = \frac{n!}{(n-j)!(j-1)!}[F(x)]^{j-1}[1-F(x)]^{n-j}f(x)$$

and the cumulative distribution function of $X_{(j)}$ is given by

$$F_{X_{(j)}}(y) = \frac{n!}{(n-j)!(j-1)!}\int_{-\infty}^{y}[F(x)]^{j-1}[1-F(x)]^{n-j}f(x)dx$$

However by definition we also have

$$\begin{aligned}F_{X_{(j)}}(y) = \mathbb{P}\{X_{(j)} \le y\} &= \mathbb{P}\{j \text{ or more of the } X_i\text{'s are less or equal than } y\} \\ &= \sum_{k=j}^{n}\binom{n}{k}[F(y)]^k[1-F(y)]^{n-k}\end{aligned}$$

So we have an interesting inequality:

$$\frac{n!}{(n-j)!(j-1)!}\int_{-\infty}^{y}[x]^{j-1}[1-x]^{n-j}dx = \sum_{k=j}^{n}\binom{n}{k}y^k[1-y]^{n-k},\ \ 0 \le y \le 1$$

Furthermore, the joint density function of the order statistics $X_{(i)}$ and $X_{(j)}$ when $1 < j$ is

$$f_{X_{(i)},X_{(j)}}(x_i,x_j) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!}[F(x_i)]^{i-1}[F(x_j)-F(x_i)]^{j-i-1}[1-F(x_j)]^{n-j}f(x_i)f(x_j).$$

Now suppose that n independent and identically distributed random variables $X_1, X_2, \cdots, X_n$ are observed, the random variable R defined by $R := X_{(n)} - X_{(1)}$ is called the range of the observed random variables. If the random variables X_i have distribution function F and density function f, then for $a \geq 0$ we have

$$\begin{aligned}\mathbb{P}\{R \leq a\} &= \mathbb{P}\{X_{(n)} - X_{(1)} \leq a\} \\ &= \iint_{x_n - x_1 \leq a} f_{X_{(1)}, X_{(n)}}(x_1, x_n) dx_1 dx_n \\ &= \int_{-\infty}^{+\infty} \int_{x_1}^{x_1 + a} \frac{n!}{(n-2)!} [F(x_n) - F(x_1)]^{n-2} f(x_1) f(x_n) dx_n dx_1\end{aligned}$$

By making $y = F(x_n) - F(x_1)$, we have

$$\int_{x_1}^{x_1 + a} [F(x_n) - F(x_1)]^{n-2} f(x_n) dx_n = \int_0^{F(x_1 + a) - F(x_1)} y^{n-2} dy = \frac{1}{n-1} [F(x_1 + a) - F(x_1)]^{n-1}$$

Thus

$$\mathbb{P}\{R \leq a\} = n \int_{-\infty}^{+\infty} [F(x_1 + a) - F(x_1)]^{n-1} f(x_1) dx_1$$

In the case when X_i's are all uniformly distributed on $(0,1)$, for $0 < a < 1$, we have

$$\begin{aligned}\mathbb{P}\{R < a\} &= n \int_0^1 [F(x_1 + a) - F(x_1)]^{n-1} f(x_1) dx_1 \\ &= n \int_0^{1-a} a^{n-1} dx_1 + n \int_{1-a}^1 (1 - x_1)^{n-1} dx_1 \\ & n(1-a)^{n-1} + a^n\end{aligned}$$

Differentiation yields the density function of the range, given in this case by

$$f_R(a) = \begin{cases} n(n-1) a^{n-2} (1-a) : 0 \leq a \leq 1 \\ 0 : \text{otherwise} \end{cases}$$

That is, the range of n independent uniform $(0,1)$ random variables is a beta random variable with parameters $n-1, 2$.

Example 3: Let (X, Y) denote a random point in the plane, and assume that the rectangular coordinates X, Y are independent standard normal random variables. We are interested in the joint distribution of R, Θ, the polar coordinate representation of (x, y)

Exchangeable Random Variables

Definition

Definition 48. *The random variables $X_1, X_2, \cdots, X_n$ are said to be exchangeable if, for every permutation $\pi(i_1, i_2, \cdots, i_n) \in S_n$, we have*

$$\mathbb{P}\{X_{i_1} \leq x_1, X_{i_2} \leq x_2, \cdots, X_{i_n} \leq x_n\} = \mathbb{P}\{X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n\}$$

That is, the n random variables are exchangeable if their joint distribution is the same no matter in which order the variables are observed.

For example, if $n = 4$ and using a particular permutation $\pi = (3,1,2,4)$, we want

$$f_{X_1,X_2,X_3,X_4}(x_1,x_2,x_3,x_4) = f_{X_3,X_1,X_2,X_4}(x_1,x_2,x_3,x_4)$$

If X_i's are discrete, this is equivalent to say

$$\mathbb{P}\{X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4\} = \mathbb{P}\{X_3 = x_1, X_1 = x_2, X_2 = x_3, X_4 = x_4\}$$

Example 1: Suppose we have an urn containing 1 red ball and 2 white balls, draw out balls one at a time and without replacement, and note the color. Define

$$X_i = \begin{cases} 1, & \text{if the } i\text{th ball is red} \\ 0, & \text{otherwise} \end{cases}$$

Then the random variables X_1, X_2, X_3 are exchangeable.

To see this, we compute the followings:

$$\mathbb{P}(X_1 = 1, X_2 = 0, X_3 = 0) = \frac{1}{3} \cdot 1 \cdot 1 = \frac{1}{3}$$

$$\mathbb{P}(X_1 = 0, X_2 = 1, X_3 = 0) = \frac{2}{3} \cdot \frac{1}{2} \cdot 1 = \frac{1}{3}$$

$$\mathbb{P}(X_1 = 0, X_2 = 0, X_3 = 1) = \frac{2}{3} \cdot \frac{1}{2} \cdot 1 = \frac{1}{3}$$

Since they are all the same, so we say X_1, X_2, X_3 are exchangeable.

Corollary

Corollary 37. *Suppose that $X_1, X_2, \cdots, X_n$ are independent and identically distributed (iid), then $X_1, X_2, \cdots, X_n$ are exchangeable.*

Proof. Let f be the probability density function for any of the X_i, then by definition we have

$$f_{X_1,X_2,\cdots,X_n}(x_1,x_2,\cdots,x_n) \overset{iid}{=} f(x_1) \cdot f(x_2) \cdots f(x_n)$$

Thus $R.H.S$ can be arranged in any order as desired. ■

Corollary

Corollary 38. *Exchangeable random variables are identically distributed.*

Proof. consider the continuous interchangeable random variables $X_1, X_2, \cdots, X_n$, and by the generalized Fubini's theorem, it's clear that

$$\begin{aligned} f_{X_i}(x_i) &= \int\int\cdots\int f_{X_1,\cdots,X_n}(x_1,\cdots x_i,\cdots,x_j,\cdots x_n)dx_1\cdots d_{x_{i-1}}d_{x_{i+1}}\cdots dx_n \\ &= \int\int\cdots\int f_{X_1,\cdots,X_n}(x_1,\cdots x_j,\cdots,x_i,\cdots x_n)dx_1\cdots d_{x_{i-1}}d_{x_{i+1}}\cdots dx_n \end{aligned}$$

Thus it also holds for any permutation $\pi \in S_n$, and thus we conclude that $f_{X_1}(x_1) = f_{X_2}(x_2) = \cdots = f_{X_n}(x_n)$ ■

Definition

Definition 49. *The random variables $X_1, X_2, \cdots$ in an infinite sequence are said to be exchangeable if the finite collection $X_1, X_2, \cdots, X_n$ are exchangeable for any finite $n \geq 1$.*

Example 2: (Pólya's Urn) Suppose we have an urn containing R_0 red balls and W_0 white balls, and $c \geq 0$ be a fixed integer. We draw a ball, note the color, replace the ball and put an additional c balls of that color in the urn as well. Rinse and repeat. Define

$$X_i = \begin{cases} 1, & \text{if the } i\text{th ball is red} \\ 0, & \text{otherwise} \end{cases}$$

Then the infinite sequence of random variables $X_1, X_2, \cdots, X_n$ are exchangeable.

To see this, we begin with an illustration that we will later generalize, note that

$$\begin{aligned} &\mathbb{P}(X_1 = 1, X_2 = 1, X_3 = 0, X_4 = 1, X_5 = 0) \\ =&\frac{R_0}{R_0+W_0} \cdot \frac{R_0+c}{R_0+W_0+c} \cdot \frac{W_0}{R_0+W_0+2c} \cdot \frac{R_0+2c}{R_0+W_0+3c} \cdot \frac{W_0+c}{R_0+W_0+4c} \end{aligned}$$

Fix any positive integer n and consider the sequence $X_1, X_2, \cdots, X_n$, we want to give an general expression for $\mathbb{P}(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n)$, from the particular case illustrated above, it is easy to see that the denominator of this more general case will be

$$(R_0+W_0)(R_0+W_0+c)(R_0+W_0+2c)\cdots(R_0+W_0+(n-1)c)$$

Note that $\sum_{i=1}^n x_i$ is the number of red balls chosen in n draws from the urn and $n - \sum_{i=1}^n x_i$ is the number of the white balls.

If $\sum x_i = n$ (all balls drawn are red), we have

$$\begin{aligned} &\mathbb{P}(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n) \\ &= \frac{R_0(R_0+c)(R_0+2c)\cdots(R_0+c(\sum x_i - 1))}{(R_0+W_0)(R_0+W_0+c)(R_0+W_0+2c)\cdots(R_0+W_0+(n-1)c)} \end{aligned}$$

and if $\sum x_i = 0$ (all balls drawn are white), we have

$$\begin{aligned}&\mathbb{P}(X_1 = x_1, X_2 = x_2, \cdots X_n = x_n)\\ &= \frac{W_0(W_0+c)(W_0+2c)\cdots(W_0+c(n-\sum x_i-1))}{(R_0+W_0)(R_0+W_0+c)(R_0+W_0+2c)\cdots(R_0+W_0+(n-1)c)}\end{aligned}$$

If $0 < \sum x_i < n$, we have

$$\begin{aligned}&\mathbb{P}(X_1 = x, X_2 = x_2, \cdots, X_n = x_n)\\ &= \frac{R_0(R_0+c)(R_0+2c)\cdots(R_0+c(\sum x_i-1))\cdot W_0(W_0+c)(W_0+2c)\cdots(W_0+c(n-\sum x_i-1))}{(R_0+W_0)(R_0+W_0+c)(R_0+W_0+2c)\cdots(R_0+W_0+(n-1)c)}\end{aligned}$$

In all cases, the probability is a function of $\sum x_i$ and not the individual positions of the 1's and 0's that make up the x_i, thus we say $X_1, X_2, \cdots, X_n$ are exchangeable.

Example 3: Let $X_1, X_2, \cdots, X_n$ be independent uniform $(0,1)$ random variables and denote their order statistics by $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$, that is $X_{(j)}$ is the j th smallest of $X_1, X_2, \cdots, X_n$. Also denote $Y_1 = X_{(1)}, Y_i = X_{(i)} - X_{(i-1)}$, show that $Y_1, Y_2, \cdots, Y_n$ are exchangeable.

Solution: The transformations yield that

$$x_i = y_1 + y_2 + \cdots + y_i$$

and the Jacobian of this transformation is just 1, so we have

$$f_{Y_1,Y_2,\cdots,Y_n}(y_1, y_2, \cdots, y_n) = f(y_1, y_1+y_2, \cdots, y_1+y_2+\cdots+y_n)$$

where f is the joint density function of the order statistics, hence we have that

$$f_{Y_1,Y_2,\cdots,Y_n}(y_1, y_2, \cdots, y_n) = n!$$

Because the preceding joint density is a symmetric function of $y_1, y_2, \cdots, y_n$, we see that the random variables are exchangeable.

Covariance, Correlation and Inequalities

Definition

Definition 50. *If random variables X,Y satisfies $\mathbb{E}\{(X-\mathbb{E}X)(Y-\mathbb{E}Y)\} < +\infty$, we then call it the covariance between X and Y and write*

$$\mathbf{Cov}(X,Y) = \mathbb{E}\{(X-\mathbb{E}X)(Y-\mathbb{E}Y)\} = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

Corollary

Corollary 39. *If X,Y are independent, then $\mathbf{Cov}(X,Y) = 0$.*

Proof. Since X,Y are independent, so for any functions h,g, we have

$$\mathbb{E}\{g(X)h(Y)\} = \mathbb{E}\{g(X)\}\mathbb{E}\{h(Y)\}$$

So we have $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$, which finishes the proof. ■

Corollary

Corollary 40. *We also have the following propositions for covariance:*

① $\mathbf{Cov}(Y,X) = \mathbf{Cov}(X,Y)$;

② $\mathbf{Cov}(X,X) = \mathbf{Var}(X)$;

③ $\mathbf{Cov}(aX,Y) = a\mathbf{Cov}(X,Y)$;

④ $\mathbf{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} \mathbf{Cov}(X_i,Y_j)$

The proofs are left as an exercise. One important remark is that we have

$$\mathbf{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathbf{Var}(X_i) + 2\sum\sum_{i<j} \mathbf{Cov}(X_i,X_j).$$

Definition

Definition 51. *The correlation of two random variables X,Y denoted by $\rho(X,Y)$, is defined as long as $\mathbf{Var}(X), \mathbf{Var}(Y)$ is positive, is defined by*

$$\rho(X,Y) = \frac{\mathbf{Cov}(X,Y)}{\sqrt{\mathbf{Var}(X)\,\mathbf{Var}(Y)}}$$

Corollary

Corollary 41.

$$-1 \leq \rho(X,Y) \leq 1$$

Proof. Suppose X,Y have variances given by σ_x^2, σ_y^2 respectively, then on the one hand,

$$\begin{aligned} 0 &\leq \mathbf{Var}\left(\frac{X}{\sigma_x} + \frac{Y}{\sigma_y}\right) \\ &= \frac{\mathbf{Var}(X)}{\sigma_x^2} + \frac{\mathbf{Var}(Y)}{\sigma_y^2} + \frac{2\,\mathbf{Cov}(X,Y)}{\sigma_x \sigma_y} \\ &= 2[1+\rho(X,Y)] \end{aligned}$$

implying that

$$-1 \leq \rho(X,Y).$$

On the other hand,

$$\begin{aligned} 0 &\leq \mathbf{Var}\left(\frac{X}{\sigma_x} - \frac{Y}{\sigma_y}\right) \\ &= \frac{\mathbf{Var}(X)}{\sigma_x^2} + \frac{\mathbf{Var}(Y)}{\sigma_y^2} - \frac{2\,\mathbf{Cov}(X,Y)}{\sigma_x \sigma_y} \\ &= 2[1-\rho(X,Y)] \end{aligned}$$

implying that

$$\rho(X,Y) \leq 1.$$

■

The correlation coefficient is a measure of the degree of linearity between X and Y. A value of $\rho(X,Y)$ near $+1$ or -1 indicates a high degree of linearity between X and Y, whereas a value near 0 indicates that such linearity is absent. A positive value of $\rho(X,Y)$ indicates that Y tends to increase when X does, whereas a negative value indicates that Y tends to decrease when X increases. If $\rho(X,Y) = 0$, then X and Y are said to be uncorrelated.

Example 1: Let $X_1, X_2, \cdots, X_n$ be independent and identically distributed random variables having variance σ^2, show that

$$\mathbf{Cov}(X_i - \overline{X}, \overline{X}) = 0$$

where $\overline{X}$ is called the sample mean, denoted by

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

Solution: We have

$$\begin{aligned}\mathbf{Cov}(X_i-\overline{X},\overline{X}) &= \mathbf{Cov}(X_i,\overline{X})-\mathbf{Cov}(\overline{X},\overline{X})\\ &= \mathbf{Cov}\left(X_i,\frac{1}{n}\sum_{j=1}^{n}X_j\right)-\mathbf{Var}(\overline{X})\\ &= \frac{1}{n}\sum_{j=1}^{n}\mathbf{Cov}(X_i,X_j)-\frac{\sigma^2}{n}\\ &= \frac{\sigma^2}{n}-\frac{\sigma^2}{n}\\ &= 0.\end{aligned}$$

Corollary

Corollary 42. *Let X,Y to be random variables, consider $X'=aX+b, Y'=cY+d$ where $a,c>0$, then*

$$\rho(X,Y)=\rho(X',Y').$$

Proof. If we denote $\mu_X=\mathbb{E}X, \mu_Y=\mathbb{E}Y$, then we have

$$\mathbb{E}X'=a\mathbb{E}X+b=a\mu_X+b$$

$$\mathbb{E}Y'=c\mathbb{E}Y+d=c\mu_Y+d$$

Also

$$\begin{aligned}\mathbf{Var}(X') &= \mathbb{E}[(X-\mu_{X'})^2]\\ &= \mathbb{E}[(aX+b-a\mu_X-b)^2]\\ &= \mathbb{E}[a^2(X-\mu_X)^2]\\ &= a^2\mathbb{E}[(X-\mu_X)^2]=a^2\sigma_X.\end{aligned}$$

Likewise, we have

$$\mathbf{Var}(Y')=c^2\sigma_Y.$$

So

$$\begin{aligned}\rho(X',Y') &= \frac{\mathbf{Cov}(X',Y')}{\sigma_{X'}\sigma_{Y'}}\\ &= \frac{\mathbb{E}[(X'-\mu_{X'})(Y'-\mu_{Y'})]}{\sigma_{X'}\sigma_{Y'}}\\ &= \frac{ac\,\mathbf{Cov}(X,Y)}{\sigma_{X'}\sigma_{Y'}}\\ &= \frac{ac\,\mathbf{Cov}(X,Y)}{(a\sigma_X)(c\sigma_Y)}\\ &= \rho(X,Y).\end{aligned}$$

■

It shows that shifting and re-scaling of random variables does not change the correlation.

Example: N people sit around a round table ($N > 5$). Each person will toss a fair coin. Anyone whose outcome is different from his/her two neighbors will receive a present. Let X be the number of people who receives presents, find $\mathbb{E}X$ and $\mathbf{Var}X$.

Solution: Number those N people from 1 to N, let X_i be the random variable defined as

$$X_i = \begin{cases} 1 : \text{if the person receives the present} \\ 0 : \text{otherwise} \end{cases}$$

then clearly $X = X_1 + X_2 + \cdots + X_n$. And observe that $\mathbb{P}(X_i = 1) = 0.25$, this is because all possible outcomes of that person and his/her neighbors are:

$$HHH, TTT, HTT, HHT, HTH, THH, THT, TTH$$

and then by linearity we have

$$\mathbb{E}X = \mathbb{E}X_1 + \cdots + \mathbb{E}X_n = \frac{N}{4}.$$

Theorem

Theorem 35. *Suppose X, Y are random variables, then*

$$\mathbb{E}|X+Y|^r \leq C_r(\mathbb{E}|X|^r + \mathbb{E}|Y|^r)$$

where $C_r = \begin{cases} 1 : 0 \leq r \leq 1 \\ 2^{r-1} : r > 1 \end{cases}$.

Proof. This follows from the fact that

$$|a+b|^r \leq C_r[|a|^r + |b|^r].$$

■

Theorem

Theorem 36. *(Hölder's Inequality)*

Let X, Y be random variables and let $p > 1, q > 1$ so that $\frac{1}{p} + \frac{1}{q} = 1$, then

$$\mathbb{E}|XY| \leq (\mathbb{E}|X|^p)^{\frac{1}{p}} (\mathbb{E}|X|^q)^{\frac{1}{q}}$$

If we take $p = q = 2$, we will get a special case of Hölder's inequality, which is known as the Cauchy-Schwarz inequality.

Theorem

Theorem 37. *Let X, Y be random variables, then*

$$\mathbb{E}|XY| \leq \sqrt{\mathbb{E}|X|^2} \cdot \sqrt{\mathbb{E}|Y|^2}$$

Corollary

Corollary 43. *Let X, Y be random variables, then*

$$|\mathbf{Cov}(X,Y)| \leq \sigma_X \cdot \sigma_Y$$

where we define

$$\sigma_X = \sqrt{\mathbf{Var}(X)}, \sigma_Y = \sqrt{\mathbf{Var}(Y)}.$$

Corollary

Corollary 44. *(Lyaponov's Inequality)*

Let X be a random variable, and $1 \leq r < s < +\infty$, then

$$(\mathbb{E}|X|^r)^{\frac{1}{r}} \leq (\mathbb{E}|X|^s)^{\frac{1}{s}}$$

Theorem

Theorem 38. *(Jensen's Inequality)*

If g is a convex function (i.e $f''(x) > 0$) and $\mathbb{E}X$ exists, then

$$g(\mathbb{E}X) \leq \mathbb{E}(g(X)).$$

Likewise, if g is concave (i.e $f''(x) < 0$) and $\mathbb{E}X$ exists, then

$$g(\mathbb{E}X) \geq \mathbb{E}(g(X)).$$

Random Vectors

When dealing with multiple random variables, it is sometimes useful to use vector and matrix notations.

Definition

Definition 52. *Let $X_1, X_2, \cdots, X_n$ be n random variables, then define the random vector $\boldsymbol{X}$ to be*

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}.$$

Here $\boldsymbol{X}$ is an n-dimensional vector because it consists of n random variables. We can further define the distribution function of the random vector $\boldsymbol{X}$ to be

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = F_{X_1,X_2,\cdots,X_n}(x_1, x_2, \cdots, x_n) = \mathbb{P}\{X_1 \leq x_1, X_2 \leq x_2, \cdots X_n \leq x_n\}.$$

If X_i's are jointly continuous, then we have the PDF written as

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = f_{X_1,X_2,\cdots,X_n}(x_1, x_2, \cdots, x_n).$$

Definition

Definition 53. *The expected value (the mean vector) of the random vector $\boldsymbol{X}$ is defined as*

$$\mathbb{E}\boldsymbol{X} = \begin{bmatrix} \mathbb{E}X_1 \\ \mathbb{E}X_2 \\ \vdots \\ \mathbb{E}X_n \end{bmatrix}.$$

Similarly, a random matrix is a matrix whose elements are random variables. In particular we can have an m by n matrix $\boldsymbol{M}$ as

$$\boldsymbol{M} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ X_{m1} & X_{m2} & \cdots & X_{mn} \end{bmatrix}.$$

Thus the mean matrix of $\boldsymbol{M}$ is given by

$$\mathbb{E}\boldsymbol{M} = \begin{bmatrix} \mathbb{E}X_{11} & \mathbb{E}X_{12} & \cdots & \mathbb{E}X_{1n} \\ \mathbb{E}X_{21} & \mathbb{E}X_{22} & \cdots & \mathbb{E}X_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbb{E}X_{m1} & \mathbb{E}X_{m2} & \cdots & \mathbb{E}X_{mn} \end{bmatrix}.$$

Corollary

Corollary 45. *(Linearity of Expectation)*

Let X be an n-dimensional random vector, $\boldsymbol{A}$ is a fixed m by n matrix and $\boldsymbol{b}$ is a fixed m-dimensional vector, define $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}$, then

$$\mathbb{E}\boldsymbol{Y} = \boldsymbol{A}\mathbb{E}\boldsymbol{X} + \boldsymbol{b}.$$

Now we would like to define the correlation and covariance matrix:

Definition

Definition 54. *(Correlation and Covariance Matrix)*

For a random vector $\boldsymbol{X}$, we define the correlation matrix $\boldsymbol{R}_X$ as

$$\boldsymbol{R}_{\boldsymbol{X}} = \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^{\boldsymbol{T}}] = \begin{bmatrix} \mathbb{E}X_1^2 & \mathbb{E}X_1X_2 & \cdots & \mathbb{E}X_1X_n \\ \mathbb{E}X_2X_1 & \mathbb{E}X_2^2 & \cdots & \mathbb{E}X_2X_n \\ \vdots & \vdots & \vdots & \vdots \\ \mathbb{E}X_nX_1 & \mathbb{E}X_nX_2 & \cdots & \mathbb{E}X_n^2 \end{bmatrix}.$$

Also, the covariance matrix $\boldsymbol{C}_X$ is defined as

$$\boldsymbol{C}_{\boldsymbol{X}} = \mathbb{E}[(\boldsymbol{X} - \mathbb{E}\boldsymbol{X})(\boldsymbol{X} - \mathbb{E}\boldsymbol{X})^{\boldsymbol{T}}] = \begin{bmatrix} \mathbf{Var}(X_1) & \mathbf{Cov}(X_1,X_2) & \cdots & \mathbf{Cov}(X_1,X_n) \\ \mathbf{Cov}(X_2,X_1) & \mathbf{Var}(X_2) & \cdots & \mathbf{Cov}(X_2,X_n) \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{Cov}(X_n,X_1) & \mathbf{Cov}(X_n,X_2) & \cdots & \mathbf{Var}(X_n) \end{bmatrix}$$

The covariance matrix is a generalization of the variance of a random variable. Remember that for a random variable, we have $\mathbf{Var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2$.

Corollary

Corollary 46. *For a random vector $\boldsymbol{X}$, we have*

$$\boldsymbol{C}_{\boldsymbol{X}} = \boldsymbol{R}_{\boldsymbol{X}} - \mathbb{E}\boldsymbol{X}\mathbb{E}\boldsymbol{X}^{\boldsymbol{T}}$$

Also, covariance matrix is symmetric, so it satisfies all the nice properties of a symmetric matrix. In particular, $\boldsymbol{C}_{\boldsymbol{X}}$ can be diagonalized and has real eigenvalues, since we assume $\boldsymbol{X}$ is a real random vector.

Definition

Definition 55. *An $n \times n$ symmetric matrix $\boldsymbol{M}$ is positive semi-definite (PSD) if for all n-dimensional vectors $\boldsymbol{b}$, we have*

$$\boldsymbol{b}^{\boldsymbol{T}} \boldsymbol{M} \boldsymbol{b} \geq 0.$$

Also, $\boldsymbol{M}$ is said to be positive definite (PD), if for all vectors $\boldsymbol{b} \neq 0$, we have

$$\boldsymbol{b}^{\boldsymbol{T}} \boldsymbol{M} \boldsymbol{b} > 0.$$

A special important property of the covariance matrix is that it is positive semi-definite (PSD).

Theorem

Theorem 39. *Let $\boldsymbol{X}$ be a random vector with n elements, then its covariance matrix $\boldsymbol{C_X}$ is positive semi-definite (PSD).*

Proof. Let $\boldsymbol{b}$ be any fixed vector with n elements. Define the random variable Y as

$$Y = \boldsymbol{b}^T (\boldsymbol{X} - \mathbb{E}\boldsymbol{X})$$

Then we have

$$\begin{aligned} 0 &\leq \mathbb{E}Y^2 \\ &= \mathbb{E}(YY^T) \\ &= \boldsymbol{b}^T \mathbb{E}[(\boldsymbol{X} - \mathbb{E}\boldsymbol{X})(\boldsymbol{X} - \mathbb{E}\boldsymbol{X})^T]\boldsymbol{b} \\ &= \boldsymbol{b}^T \boldsymbol{C_X} \boldsymbol{b}. \end{aligned}$$

■

Theorem

Theorem 40. *Let $\boldsymbol{X}$ be a random vector with n elements, then its covariance matrix $\boldsymbol{C_X}$ is positive definite (PD) if and only if all its eigenvalues are larger than zero. Equivalently, $\boldsymbol{C_X}$ is positive definite (PD) if and only if* $\det(\boldsymbol{C_X}) > 0$.

Functions of Random Vectors

Theorem

Theorem 41. *Let* $\boldsymbol{X} = \begin{bmatrix} X_1 & X_2 & \cdots & X_n \end{bmatrix}^T$ *be an n-dimensional random vector with joint PDF* $f_{\boldsymbol{X}}(\boldsymbol{x})$*. Let* $G : \mathbb{R}^n \to \mathbb{R}^n$ *be a continuous and invertible function with continuous partial derivatives and let* $H = G^{-1}$*. Suppose a random vector* $\boldsymbol{Y} = \begin{bmatrix} Y_1 & Y_2 & \cdots & Y_n \end{bmatrix}^T$ *is given by* $\boldsymbol{Y} = G(\boldsymbol{X})$*, hence* $\boldsymbol{X} = H(\boldsymbol{Y})$*, and*

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} H_1(Y_1, Y_2, \cdots, Y_n) \\ H_2(Y_1, Y_2, \cdots, Y_n) \\ \vdots \\ H_n(Y_1, Y_2, \cdots, Y_n) \end{bmatrix}$$

Then the PDF of $\boldsymbol{Y}$ *is given by*

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = f_{\boldsymbol{X}}(H(\boldsymbol{y}))|J|$$

where J is the Jacobian of H defined by

$$J = \begin{bmatrix} \frac{\partial H_1}{\partial y_1} & \frac{\partial H_1}{\partial y_2} & \cdots & \frac{\partial H_1}{\partial y_n} \\ \frac{\partial H_2}{\partial y_1} & \frac{\partial H_2}{\partial y_2} & \cdots & \frac{\partial H_2}{\partial y_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial H_n}{\partial y_1} & \frac{\partial H_n}{\partial y_2} & \cdots & \frac{\partial H_n}{\partial y_n} \end{bmatrix}.$$

Example 1: Suppose $Y_1, Y_2 \sim N(0,1)$ are independent random variables, find the probability density function of $\sqrt{Y_1^2 + Y_2^2}$.

Solution: We first define two new variables:

$$U = \sqrt{Y_1^2 + Y_2^2}, V = Y_2$$

Then we have

$$Y_1 = \pm\sqrt{U^2 - V^2}; Y_2 = V$$

Thus we may define two linear maps given by

$$\Phi_1 : (+\sqrt{u^2 - v^2}, v) \mapsto (u, v)$$

$$\Phi_2 : (-\sqrt{u^2 - v^2}, v) \mapsto (u, v)$$

The Jacobians are given by

$$\mathscr{J}_1 = \det \begin{bmatrix} u(\sqrt{u^2 - v^2})^{-1} & -v(\sqrt{u^2 - v^2})^{-1} \\ 0 & 1 \end{bmatrix} = u(\sqrt{u^2 - v^2})^{-1}$$

$$\mathscr{J}_1 = \det \begin{bmatrix} -u(\sqrt{u^2-v^2})^{-1} & v(\sqrt{u^2-v^2})^{-1} \\ 0 & 1 \end{bmatrix} = -u(\sqrt{u^2-v^2})^{-1}$$

Then we have

$$\begin{aligned} f_{U,V}(u,v) &= f_{Y_1,Y_2}(\sqrt{u^2-v^2},v)|\mathscr{J}_1| + f_{Y_1,Y_2}(-\sqrt{u^2-v^2},v)|\mathscr{J}_2| \\ &= \left(\frac{u}{\sqrt{u^2-v^2}}\right)\left[\frac{1}{2\pi}e^{-\frac{u^2}{2}} + \frac{1}{2\pi}e^{-\frac{u^2}{2}}\right]. \end{aligned}$$

And we have

$$f_{U,V}(u,v) = \begin{cases} \dfrac{ue^{-\frac{u^2}{2}}}{\pi\sqrt{u^2-v^2}} : |v| < u \\ \\ \text{DOES NOT EXIST} : \text{otherwise} \end{cases}$$

Now we will find the distribution restricted to U only, and we have

$$\begin{aligned} f_U(u) &= \int_{-\infty}^{+\infty} \frac{ue^{-\frac{u^2}{2}}}{\pi\sqrt{u^2-v^2}}dv \\ &= \frac{ue^{\frac{u^2}{2}}}{\pi}\int_{-u}^{u}\frac{1}{\sqrt{u^2-v^2}}dv \\ &= \frac{1}{\pi}e^{-\frac{u^2}{2}}\int_{-\frac{\pi}{2}}^{+\frac{\pi}{2}}\frac{u\cos(\theta)d\theta}{\sqrt{1-\sin^2(\theta)}} \\ &= ue^{-\frac{u^2}{2}}, u \geq 0 \end{aligned}$$

So we have

$$f_U(u) = \begin{cases} ue^{-\frac{u^2}{2}} : u \geq 0 \\ 0 : \text{otherwise} \end{cases}.$$

Example 2: Suppose $X_1, X_2, X_3 \sim N(0,1)$ are independent standard normal random variables, and let $Y_1 = X_1 + X_2 + X_3, Y_2 = X_1 - X_2, Y_3 = X_1 - X_3$, find the joint density function of Y_1, Y_2, Y_3.

Solution : We have

$$X_1 = \frac{Y_1+Y_2+Y_3}{3}, X_2 = \frac{Y_1-2Y_2+Y_3}{3}, X_3 = \frac{Y_1+Y_2-2Y_3}{3}$$

and thus the Jacobian of the transformation is given by

$$\mathscr{J} := \det \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & -\frac{2}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & -\frac{2}{3} \end{bmatrix}$$

and we have

$$\begin{aligned} f_{Y_1,Y_2,Y_3}(y_1,y_2,y_3) &= f_{X_1,X_2,X_3}\left(\frac{Y_1+Y_2+Y_3}{3}, \frac{Y_1-2Y_2+Y_3}{3}, \frac{Y_1+Y_2-2Y_3}{3}\right)\cdot\frac{1}{3} \\ &= \frac{1}{3(2\pi)^{3/2}}e^{-Q(y_1,y_2,y_3)/2} \end{aligned}$$

where

$$Q(y_1,y_2,y_3)=\frac{y_1^2}{3}+\frac{2}{3}y_2^2+\frac{2}{3}y_3^2-\frac{2}{3}y_2y_3.$$

Example: Suppose U_1,U_2 are random variables uniformly distributed on $(0,1)$, find the distribution function for X_1 and X_2 where $X_1=\sin(2\pi U_2)\sqrt{-2\ln(U_1)}, X_2=\cos(2\pi U_2)\sqrt{-2\ln(U_1)}$.

Solution : We first find the "inverse" Jacobian:

$$\mathscr{J}^{-1}=\det\begin{pmatrix}\frac{\partial X_1}{\partial U_1} & \frac{\partial X_1}{\partial U_2}\\ \\ \frac{\partial X_2}{\partial U_1} & \frac{\partial X_2}{\partial U_2}\end{pmatrix}$$

$$=\det\begin{pmatrix}-\dfrac{\sin(2\pi U_2)}{U_1\sqrt{-2\ln(U_1)}} & 2\pi\cos(2\pi U_2)\sqrt{-2\ln(U_1)}\\ \\ -\dfrac{\cos(2\pi U_2)}{U_1\sqrt{-2\ln(U-1)}} & -2\pi\sin(2\pi U_2)\sqrt{-2\ln(U_1)}\end{pmatrix}$$

$$=\frac{2\pi\sin^2(2\pi u_2)}{U_1}+\frac{2\pi\cos^2(2\pi U_2)}{U_1}$$
$$=\frac{2\pi}{U_1}$$

So the desired Jacobian is given by

$$\mathscr{J}=\frac{1}{\mathscr{J}^{-1}}=\frac{U_1}{2\pi}$$

Now according to the relations given, solve for U_1,U_2 in terms of X_1,X_2, we get:

$$U_1=e^{-\frac{X_1^2+X_2^2}{2}}, U_2=\frac{\arctan(X_1/X_2)}{2\pi}$$

Now, we have

$$f_{X_1,X_2}(x_1,x_2)=f_{U_1,U_2}(u_1(x_1,x_2),u_2(x_1,x_2))\cdot\left|\frac{U_1}{2\pi}\right|$$

where

$$f_{U_1,U_2}(u_1,u_2)=\begin{cases}1:(u_1,u_2)\in(0,1)^2\\ 0:\text{otherwise}\end{cases},$$

so we actually have

$$f_{X_1,X_2}(x_1,x_2)=1\cdot\frac{U_1}{2\pi}=\frac{1}{2\pi}e^{-\frac{x_1^2+x_2^2}{2}}$$
$$=\left(\frac{1}{\sqrt{2\pi}}e^{-x_1^2/2}\right)\left(\frac{1}{\sqrt{2\pi}}e^{-x_2^2/2}\right).$$

And thus we have $X_1,X_2\sim N(0,1)$, the standard normal distribution.

Example: Let (X,Y) denote a random point in the plane, and the joint pdf is given by

$$f_{X,Y}(x,y) = \begin{cases} 1 : x^2 + y^2 \leq 1 \\ 0 : \text{otherwise} \end{cases}$$

Chapter 5

Limit Theorems

Limit Theorems

5.1.1 Law of Large Numbers

The law of large numbers has a very central role in probability and statistics. It states that if you repeat an experiment independently a large number of times and average the result, what you obtain should be close to the expected value. There are two main versions of the law of large numbers. They are called the weak and strong laws of the large numbers. The difference between them is mostly theoretical. In this section, we state and prove the weak law of large numbers (WLLN). Before discussing the WLLN, let us define the sample mean.

Definition

Definition 56. *For i.i.d random variables $X_1, X_2, \cdots, X_n$, the sample mean, denoted by $\overline{X}$, is defined as*

$$\overline{X} := \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

Another common notation for the sample mean is M_n, if the X_i's has CDF $F_X(x)$, we might show the sample mean by $M_n(X)$ to indicate the distribution of the X_i's.

The sample mean, $\overline{X} = M_n(X)$ is also a random variable and we have

$$\mathbb{E}[\overline{X}] = \mathbb{E}[X] \qquad\qquad \mathbf{Var}(\overline{X}) = \frac{\mathbf{Var}(X)}{n}$$

To see this, we have

$$\mathbb{E}(\overline{X}) = \frac{\mathbb{E}X_1 + \mathbb{E}X_2 + \cdots + \mathbb{E}X_n}{n} = \frac{n\mathbb{E}X}{n} = \mathbb{E}X.$$

and

$$\begin{aligned}\mathbf{Var}(\overline{X}) &= \frac{\mathbf{Var}(X_1+X_2+\cdots+X_n)}{n^2} && \text{(since } \mathbf{Var}(aX)=a^2\,\mathbf{Var}(X)\text{)}\\ &= \frac{\mathbf{Var}(X_1)+\cdots+\mathbf{Var}(X_n)}{n^2}\\ &= \frac{n\,\mathbf{Var}(x)}{n^2}\\ &= \frac{\mathbf{Var}(X)}{n}.\end{aligned}$$

Theorem

Theorem 42. *(The Weak Law Of Large Numbers (WLLN) (aka Khinchin's Theorem)*

Let $X_1, X_2, \cdots, X_n$ be i.i.d random variables with a finite expected value $\mathbb{E}X_i = \mu < +\infty$, then for any $\varepsilon > 0$,

$$\lim_{n\to+\infty} \mathbb{P}\{|\overline{X} - \mu| \geq \varepsilon\} = 0.$$

Proof. If $\mathbf{Var}(X) = \sigma^2 < +\infty$, then we may apply Chebyshev's inequality directly:

$$\mathbb{P}\{|\overline{X} - \mu| \geq \varepsilon\} \leq \frac{\mathbf{Var}(\overline{X})}{\varepsilon^2} = \frac{\mathbf{Var}(X)}{n\varepsilon^2} \to 0, \text{as } n \to +\infty.$$

■

Theorem

Theorem 43. *(The Strong Law of Large Numbers (SLLN) (aka Kolmogorov's Law))*

Let $X_1, X_2, \cdots, X_n$ be i.i.d random variables with finite expected value, then $\overline{X_n} \xrightarrow{a.s} \mu$ when $n \to +\infty$, i.e

$$\mathbb{P}\{\lim_{n\to+\infty} \overline{X}_n = \mu\} = 1.$$

Corollary

Corollary 47. *Let $\{X_n\}$ be a sequence of i.i.d random variables with finite expected value, and $\mathbb{E}X = \mu$ to be the mean, $\mathbf{Var}(X) = \sigma^2$ to be the variance. We define the sample mean $\overline{X}_n$ and sample variance S_n^2 to be*

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

Then we have

$$S_n^2 \xrightarrow{P} \sigma^2.$$

Proof. We first observe that

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

According to the law of large numbers, we have

$$\frac{1}{n}\sum_{i=1}^{n}\overline{X}_i^2 \xrightarrow{P} \mathbb{E}(X^2)$$

Also given that $\mathbb{E}X^2 = \sigma^2 + \mu^2$, thus

$$\frac{1}{n}\sum_{i=1}^{n}X_i^2 \xrightarrow{P} \sigma^2 + \mu^2.$$

Also, by *LLN*, $\overline{X}_n \xrightarrow{P} \mu$, and by continuous mapping theorem, we have

$$\overline{X}_n^2 \xrightarrow{P} \mu^2.$$

So we have

$$\frac{1}{n}\sum_{i=1}^{n}X_i^2 \xrightarrow{P} \sigma^2 + \mu^2 \quad \text{and} \quad \overline{X}_n^2 \xrightarrow{P} \mu^2$$

Therefore

$$S_n^2 \xrightarrow{P} (\sigma^2 + \mu^2) - \mu^2 = \sigma^2, \text{as } n \to +\infty.$$

■

Central Limit Theorem

Theorem

Theorem 44. *(Lindeberg–Lévy Central Limit Theorem)*

Let $X_1, X_2, \cdots, X_n$ be i.i.d random variables with expected value $\mathbb{E}X_i = \mu < +\infty$ and variance $0 < \mathbf{Var}(X_i) = \sigma^2 < +\infty$, then the random variable

$$Z_n = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$$

converges in distribution to the standard normal random variable as n goes to infinity, that is

$$\lim_{n \to +\infty} \mathbb{P}\{Z_n \leq x\} = \Phi(x), \quad \forall x \in \mathbb{R}$$

where $\Phi(x)$ is the standard normal CDF.

Proof. Assume $\{X_1, X_2, \cdots\}$ are i.i.d random variables with mean μ and finite variance σ^2, then the sum $X_1 + \cdots + X_n$ has mean $n\mu$ and variance $n\sigma^2$, consider

$$Z_n = \frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sqrt{n\sigma^2}} = \sum_{i=1}^{n} \frac{X_i - \mu}{\sqrt{n\sigma^2}} = \sum_{i=1}^{n} \frac{1}{\sqrt{n}} Y_i$$

where we define $Y_i = \frac{X_i - \mu}{\sigma}$, each with 0 mean and unit variance ($\mathbf{Var}(Y) = 1$). The characteristic function of Z_n is given by

$$\varphi_{Z_n}(t) = \left[\varphi_{Y_1}\left(\frac{t}{\sqrt{n}} \right) \right]^n$$

Given the fact that they are i.i.d. Then the characteristic function of Y_1 is, by Taylor's theorem,

$$\varphi_{Y_1}\left(\frac{t}{\sqrt{n}} \right) = 1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n} \right), \quad \left(\frac{t}{\sqrt{n}} \right) \to 0.$$

where $o(\cdot)$ is the little o notation for some function t that goes to zero more rapidly than $\frac{t^2}{n}$ does. Also since

$$e^x = \lim_{n \to +\infty} \left(1 + \frac{x}{n} \right)^n$$

we then have

$$\varphi_{Z_n}(t) = \left(1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n} \right) \right)^n \to e^{-\frac{1}{2}t^2}, \quad n \to +\infty$$

All of the higher terms vanish in the limit $n \to +\infty$, the right hand side equals the characteristic function of a standard normal distribution $N(0,1)$. ■

An interesting fact about CLT is that it does not matter what the distribution is. To get a feeling of the CLT, let's first assume that X_i's are $Bernoulli(p)$, then $EX_i = p$, $\mathbf{Var}(X_i) = p(1-p)$, also $Y_n = X_1 + \cdots + X_n$ has $Binomial(n,p)$ distributions, thus

$$Z_n = \frac{Y_n - np}{\sqrt{np(1-p)}}$$

The figure below (5.1) shoes the PMF of Z_n for different values of n, the shape of the PMF gets closer to a normal PDF as n increases.

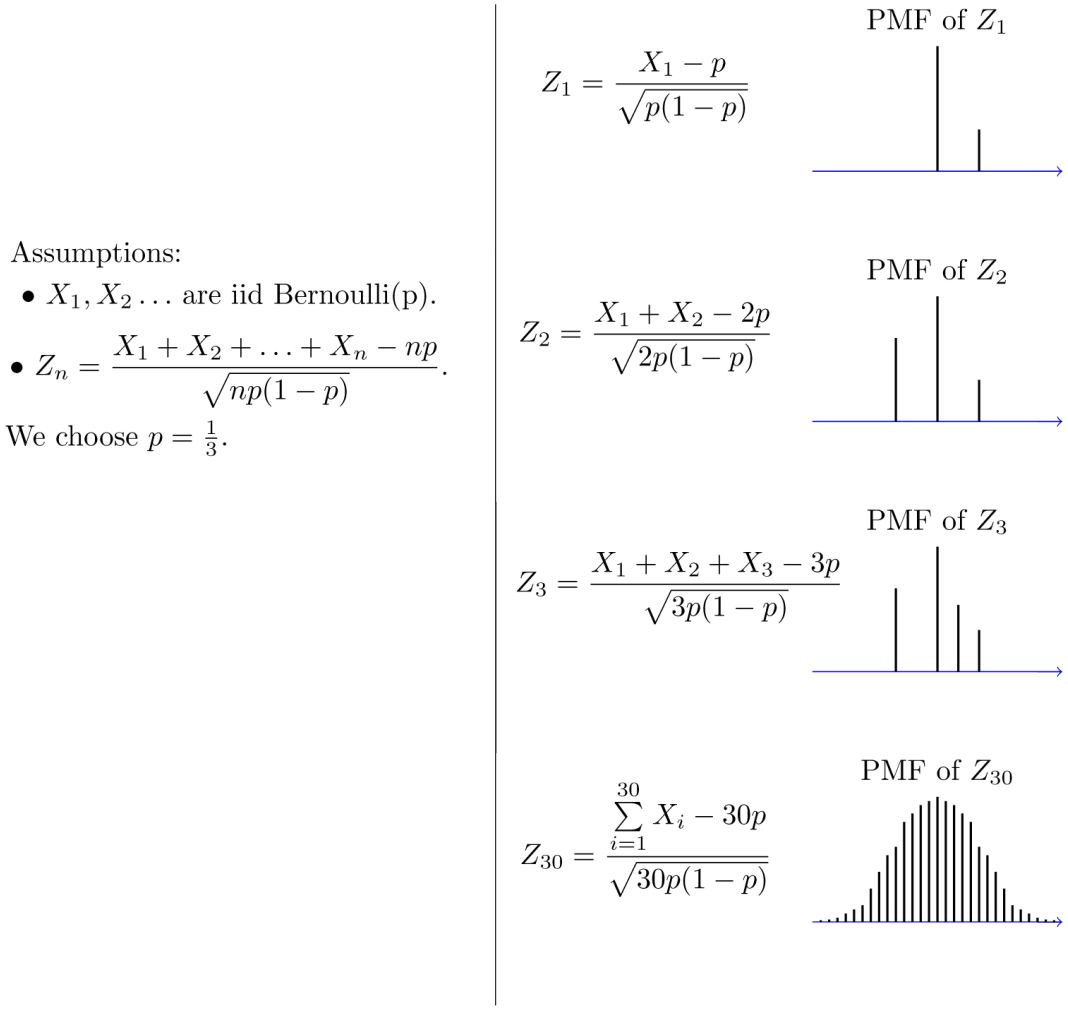


Figure 5.1: Z_n is the normalized sum of n independent $Bernoulli(p)$ random variables. The shape of its PMF, $P_{Z_n}(z)$, resembles the normal curve as n increases.

As another example, let's assume that X_i's are $Uniform(0,1)$, then $EX_i = \frac{1}{2}$, $\mathbf{Var}(X_i) = \frac{1}{12}$, in this case

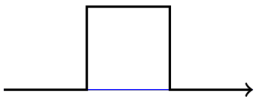
$$Z_n = \frac{X_1 + \cdots + X_n - \frac{n}{2}}{\sqrt{n/12}}$$

Assumptions:

- $X_1, X_2 \ldots$ are iid Uniform(0,1).
- $Z_n = \dfrac{X_1 + X_2 + \ldots + X_n - \frac{n}{2}}{\sqrt{\frac{n}{12}}}$.

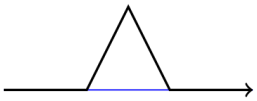
$$Z_1 = \frac{X_1 - \frac{1}{2}}{\sqrt{\frac{1}{12}}}$$

PDF of Z_1



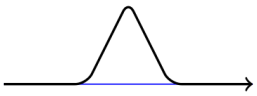
$$Z_2 = \frac{X_1 + X_2 - 1}{\sqrt{\frac{2}{12}}}$$

PDF of Z_2



$$Z_3 = \frac{X_1 + X_2 + X_3 - \frac{3}{2}}{\sqrt{\frac{3}{12}}}$$

PDF of Z_3



$$Z_{30} = \frac{\sum_{i=1}^{30} X_i - \frac{30}{2}}{\sqrt{\frac{30}{12}}}$$

PDF of Z_{30}

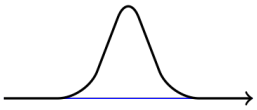


Figure 5.2: Z_n is the normalized sum of n independent $Uniform(0,1)$ random variables. The shape of its PDF, $f_{Z_n}(z)$, gets closer to the normal curve as n increases.

The importance of the central limit theorem stems from the fact that, in many real applications, a certain random variable of interest is a sum of a large number of independent random variables. In these situations, we are often able to use the CLT to justify using the normal distribution. Examples of such random variables are found in almost every discipline. Here are a few:

① Laboratory measurement errors are usually modeled by normal random variables;

② In communication and signal processing, Gaussian noise is the most frequently used model for noise;

③ In finance, the percentage changes in the prices of some assets are sometimes modeled by normal random variables;

④ When we do random sampling from a population to obtain statistical knowledge about the population, we often model the resulting quantity as a normal random variable.

Here is the steps that we need in order to apply the CLT:

HOW TO APPLY CLT

(i) : Write the random variable of interest, Y as the sum of n i.i.d random variables X_i's:

$$Y = X_1 + X_2 + \cdots + X_n.$$

(ii) : Find $\mathbb{E}Y$ and $\mathbf{Var}(Y)$ by noting that

$$\mathbb{E}Y = n\mu, \mathbf{Var}(Y) = n\sigma^2$$

where $\mu = \mathbb{E}X_i$ and $\sigma^2 = \mathbf{Var}(X_i)$.

(iii) : According to CLT, conclude that

$$\frac{Y - \mathbb{E}Y}{\sqrt{\mathbf{Var}(Y)}} = \frac{Y - n\mu}{\sigma\sqrt{n}}$$

is approximately standard normal, thus to find $\mathbb{P}\{y_1 \leq Y \leq Y_2\}$, we can write

$$\begin{aligned}\mathbb{P}\{y_1 \leq Y \leq Y_2\} &= \mathbb{P}\left(\frac{y_1 - n\mu}{\sigma\sqrt{n}} \leq \frac{Y - n\mu}{\sigma\sqrt{n}} \leq \frac{y_2 - n\mu}{\sigma\sqrt{n}}\right) \\ &\approx \Phi\left(\frac{y_2 - n\mu}{\sigma\sqrt{n}}\right) - \Phi\left(\frac{y_1 - n\mu}{\sigma\sqrt{n}}\right).\end{aligned}$$

where Φ is the standard normal CDF given by

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy$$

Example : A bank teller serves customers standing in the queue one by one. Suppose that the service time X_i for customer i has mean $\mathbb{E}X_i = 2$ (minutes) and $\mathbf{Var}(X_i) = 1$. We assume that service times for different bank customers are independent. Let Y be the total time the bank teller spends serving 50 customers. Find $\mathbb{P}(90 < Y < 110)$.

Solution : Following the steps mentioned above, let

$$Y = X_1 + X_2 + \cdots + X_n$$

where $n = 50, \mathbb{E}X_i = \mu = 2, \mathbf{Var}(X_i) = \sigma^2 = 1$, thus we can write

$$\begin{aligned}\mathbb{P}\{90 < Y < 110\} &= \mathbb{P}\left\{\frac{90 - n\mu}{\sigma\sqrt{n}} < \frac{Y - n\mu}{\sigma\sqrt{n}} < \frac{110 - n\mu}{\sigma\sqrt{n}}\right\} \\ &= \mathbb{P}\left\{\frac{90 - 100}{\sqrt{50}} < \frac{Y - n\mu}{\sigma\sqrt{n}} < \frac{110 - 100}{\sqrt{50}}\right\} \\ &= \mathbb{P}\left\{-\sqrt{2} < \frac{Y - n\mu}{\sigma\sqrt{n}} < \sqrt{2}\right\}\end{aligned}$$

By CLT, $\frac{Y-n\mu}{\sigma\sqrt{n}}$ is approximately standard normal, so we can write

$$\mathbb{P}(90 < Y < 110) \approx \Phi(\sqrt{2}) - \Phi(-\sqrt{2}) = 0.8427.$$

Example : In a communication system each data packet consists of 1000 bits. Due to the noise, each bit may be received in error with probability 0.1 . It is assumed bit errors occur independently. Find the probability that there are more than 120 errors in a certain data packet.

Solution: Let's define X_i as the indicator random variable for the ith bit in the packet. That is, $X_i = 1$ if the ith bit is received in error, and $X_i = 0$ otherwise. Then the X_i's are i.i.d. and $X_i \sim Bernoulli(p = 0.1)$. If Y is the total number of bit errors in the pocket, we have

$$Y = X_1 + X_2 + \cdots + X_n$$

Since $X_i \sim Bernoulli(p = 0.1)$, then we have

$$\mathbb{E}X_i = \mu = p = 0.1; \mathbf{Var}(X_i) = \sigma^2 = p(1-p) = 0.09$$

By CLT, we have

$$\begin{aligned}
\mathbb{P}\{Y > 120\} &= \mathbb{P}\left\{\frac{Y-n\mu}{\sigma\sqrt{n}} > \frac{120-n\mu}{\sigma\sqrt{n}}\right\} \\
&= \mathbb{P}\left\{\frac{Y-n\mu}{\sigma\sqrt{n}} > \frac{120-100}{90}\right\} \\
&\approx 1 - \Phi\left(\frac{20}{\sqrt{90}}\right) \\
&= 0.0175.
\end{aligned}$$

In general, if σ is not easy to calculate, we can approximate σ by S_n, which we introduced earlier. We have

$$S_n \xrightarrow{P} \mu.$$

So we have

$$\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} = \left(\frac{\sigma}{S_n}\right)\left(\frac{\overline{X_n} - \mu}{\sigma/\sqrt{n}}\right).$$

Modes of Convergence

Definition

Definition 57. *Let $\{F_n\}$ be a sequence of distribution functions, if there exists a distribution function F such that*

$$\lim_{n\to+\infty} F_n(x) = F(x)$$

at every point x which F is continuous, we say that F_n converges in law (or weakly) to F, and we write $F_n \xrightarrow{\omega} F$;

Let $\{X_n\}$ be a sequence of random variables and $\{F_n\}$ be the corresponding distribution functions, we say that X_n converges in distribution (or law) to X if there exists an random variable X with distribution function F such that $F_n \xrightarrow{\omega} F$. We write $X_n \xrightarrow{L} X$.

Theorem

Theorem 45. *Let $\{X_n\}, X$ be continuous random variables such that $\lim_{n\to+\infty} f_n(x) = f(x)$ for a.e $x \in \mathbb{R}$ where f_n, f are the probability density functions of X_n and X respectively, then $X_n \xrightarrow{L} X$.*

Definition

Definition 58. *Let $\{X_n\}$ be a sequence of random variables defined on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$, we say that the sequence $\{X_n\}$ converges in probability to the random variable X if $\forall \varepsilon > 0$,*

$$\lim_{n\to+\infty} \mathbb{P}\{|X_n - X| > \varepsilon\} = 0$$

And we denote by $X_n \xrightarrow{P} X$.

Example : Let X be a random variable and $X_n = X + Y_n$, where

$$EY_n = \frac{1}{n} \quad \mathbf{Var}(Y_n) = \frac{\sigma^2}{n}$$

where $\sigma > 0$ is a constant. Show that $X_n \xrightarrow{P} X$.

Solution:
By triangle inequality, we have

$$|Y_n| \leq |Y_n - \mathbb{E}Y_n| + |\mathbb{E}Y_n| = |Y_n - \mathbb{E}Y_n| + \frac{1}{n}.$$

Then, $\forall \varepsilon > 0$, we have

$$\begin{aligned}
\mathbb{P}\{|X_n - X| \geq \varepsilon\} &= \mathbb{P}\{|Y_n| \geq \varepsilon\} \\
&\leq \mathbb{P}\left\{|Y_n - \mathbb{E}Y_n| + \frac{1}{n} \geq \varepsilon\right\} \\
&= \mathbb{P}\left\{|Y_n - \mathbb{E}Y_n| \geq \varepsilon - \frac{1}{n}\right\} \\
&\leq \frac{\mathbf{Var}(Y_n)}{\left(\varepsilon - \frac{1}{n}\right)^2} \qquad \text{By Chebyshev's Inequality} \\
&= \frac{\sigma^2}{n\left(\varepsilon - \frac{1}{n}\right)^2} \to 0, \qquad \text{as } n \to +\infty.
\end{aligned}$$

Therefore we conclude that $X_n \xrightarrow{P} X$.

As we mentioned previously, convergence in probability is stronger than convergence in distribution. That is, if $X_n \xrightarrow{P} X$ then $X_n \xrightarrow{d} X$. However the converse is not necessarily true. To see this, let $X_1, X_2, \cdots, X_3$ be a sequence of i.i.d $Bernoulli(0.5)$ random variables, Let also $X \sim Bernoulli(0.5)$ be independent from the X_i's, then $X_n \to X$, however X_n does not converge in probability to X since $|X_n - X|$ is in fact also a $Bernoulli(0.5)$ random variable and

$$\mathbb{P}\{|X_n - X| \geq \varepsilon\} = 0.5$$

A special case in which the converse true is when $X_n \xrightarrow{d} c$ where c is a constant. In this case convergence in distribution implies convergence in probability.

Theorem

Theorem 46. *If $X_n \xrightarrow{d} c$ where c is a constant, then $X_n \xrightarrow{P} c$.*

Proof. Suppose X_n has distribution function F_X, then by definition we know that $\lim_{n\to+\infty} F_X = c$. Now for anny $k > 0$ we have

$$\begin{aligned}
\mathbb{P}\{|X_n - c| < k\} &= \mathbb{P}\{-k < X_n - c < k\} \\
&= \mathbb{P}\{-k + c < X_n < k + c\} \\
&= F_X(k + c) - F_X(-k + c) \\
&= 1
\end{aligned}$$

thus $X_n \xrightarrow{P} c$. ■

Corollary

Corollary 48. ① $X_n \xrightarrow{P} X \iff X_n - X \xrightarrow{P} 0$;

② *If $X_n \xrightarrow{P} X$ and $X_n \xrightarrow{P} Y$, then $\mathbb{P}\{X = Y\} = 1$.*

Corollary

Corollary 49. *If $X_n \xrightarrow{P} X$, then $X_n - X_m \xrightarrow{P} 0$.*

Corollary

Corollary 50. *If $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$, then*

$$X_n \pm Y_n \xrightarrow{P} X \pm Y$$

Theorem

Theorem 47. *(Continuous Mapping Theorem)*

Let $X_n \xrightarrow{P} X$ and g is a continuous function defined on $\mathbb{R}$, then $g(X_n) \xrightarrow{P} g(X)$.

Proof. ■

Theorem

Theorem 48. *If $X_n \xrightarrow{P} X$, then $X_n \xrightarrow{L} X$.*

In general, the reverse of this theorem is not true!

Proof. ■

Theorem

Theorem 49. *(Slutsky's Theorem)*

If $|X_n - Y_n| \xrightarrow{P} 0$ and $Y_n \xrightarrow{P} Y$, then $X_n \xrightarrow{P} Y$.

Proof. ■

Theorem

Theorem 50. *(Cramer's Theorem)*

If $X_n \xrightarrow{L} X, Y_n \xrightarrow{P} c$ where c is a constant, then

(i) : $X_n \pm Y_n \xrightarrow{L} X \pm c$;

(ii) : $X_n Y_n \xrightarrow{L} cX$ if $c \neq 0$ and $X_n Y_n \xrightarrow{P} 0$ if $c = 0$;

(iii) : $\dfrac{X_n}{Y_n} \xrightarrow{L} \dfrac{X}{c}$ if $c \neq 0$.

Example: Let $X_1, X_2, \cdots$ be a sequence of *i.i.d* random variables with common probability density function $f(x) = e^{-(x-\theta)}$ for $x \geq \theta$ and $= 0$ otherwise. Let $\overline{X}_n$ denote the sample mean given by $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

(a) Show that $\overline{X}_n \xrightarrow{P} 1 + \theta$;

(b) Show that for any sequence $Y_1, Y_2, \cdots, Y_n$ of random variables and a constant c, if $Y_n \xrightarrow{L} c$, then $Y_n \xrightarrow{P} c$;

(c) Show that $\min(X_1, X_2, \cdots, X_n) \xrightarrow{P} \theta$.

Solution:

(a) : First we find $\mathbb{E}X$, since those are *i.i.d*, they must have the same $\mathbb{E}X$ value:

$$\begin{aligned}
\mathbb{E}X &= \int_{-\infty}^{+\infty} x f_X(x) dx \\
&= \int_{\theta}^{+\infty} x e^{-(x-\theta)} dx \\
&= -x e^{-(x-\theta)} \Big|_{\theta}^{+\infty} - \int_{\theta}^{+\infty} -e^{-(x-\theta)} dx \\
&= \theta + 1.
\end{aligned}$$

Then, we find out that what we want to show is just $\overline{X}_n \xrightarrow{P} \mathbb{E}X$, to further prove this we consider for any $\varepsilon > 0$, the term $\mathbb{P}\{|\overline{X}_n - \mathbb{E}X| > \varepsilon\}$, and by Chevbeshev's inequality,

$$\begin{aligned}
\mathbb{P}\{|\overline{X}_n - \mathbb{E}X| > \varepsilon\} &\leq \frac{\mathbf{Var}(\overline{X}_n)}{\varepsilon^2} \\
&= \frac{\frac{1}{n}\mathbf{Var}(X)}{\varepsilon^2} \\
&= \frac{\mathbf{Var}(X)}{\varepsilon^2} \cdot \frac{1}{n} \to 0, \text{as } n \to +\infty.
\end{aligned}$$

So we have proved that $\overline{X}_n \xrightarrow{P} 1 + \theta$.

(b) : By definition, $Y_n \xrightarrow{L} c$ implies $\lim_{n\to+\infty} F_{Y_n} = c$ where c is a constant and F_{Y_n} is the distribution function of the random variable Y_n, and we have

$$\begin{aligned}
\mathbb{P}\{|Y_n - c| < \varepsilon\} &= \mathbb{P}\{-\varepsilon < Y_n - c < \varepsilon\} \\
&= \mathbb{P}\{-\varepsilon + c < Y_n < \varepsilon + c\} \\
&= F_{Y_n}(\varepsilon + c) - F_{Y_n}(-\varepsilon + c) \\
&\overset{\text{as } n \to +\infty}{\to} 1
\end{aligned}$$

Thus we have proved that $Y_n \xrightarrow{P} c$.

(c) : Define $X_0 = \min(X_1, X_2, \cdots, X_n)$, then by definition,

$$\mathbb{P}(X_0 \geq x) \implies \mathbb{P}(X_1 \geq x, X_2 \geq x, \cdots, X_n \geq x) = \left(\int_x^{+\infty} e^{-(t-\theta)} dt \right)^n$$
$$= e^{-n(x-\theta)}, \text{if } x > \theta.$$

Thus we have

$$f_{Y_0}(x) = \begin{cases} 1 : x \leq \theta \\ e^{-n(y-\theta)} : y > \theta \end{cases}$$

and the distribution function is

$$F_{Y_0}(x) = \begin{cases} 0 : x \leq \theta \\ e^{-n(x-\theta)} : x > \theta \end{cases}$$

as $n \to +\infty$, $F_{Y_0}(x) \to \begin{cases} 0 : x \leq \theta \\ 1 : y > \theta \end{cases}$ and hence we have $Y_n \xrightarrow{L} \theta$ where θ is a constant, and thus by (b), $Y_n \xrightarrow{P} \theta$.

Example: Let $X_1, X_2, \cdots, X_n$ be *i.i.d* random variables with common mean 0 and variance 1, suppose $\beta_4 = \mathbb{E}X^4 < +\infty$. Find the limiting distribution of the random variable

$$W_n = \sqrt{n} \frac{\sum_{i=1}^n X_i}{\sum_{i=1}^n X_i^2}.$$

Solution: Note that

$$W_n = \sqrt{n} \frac{\sum_{i=1}^n X_i}{\sum_{i=1}^n X_i^2} = \frac{\frac{\sum X_i}{\sqrt{n}}}{\frac{\sum X_i^2}{n}}$$

We define

$$U_n = \frac{\sum_{i=1}^n X_i}{\sqrt{n}}, V_n = \frac{\sum_{i=1}^n X_i^2}{n}$$

Then, according to the properties of X_i, we know that $\mu = 0, \sigma = 1$, and the Central Limit Theorem states that

$$\frac{\sum_{i=1}^n X_n - \mu}{\sigma / \sqrt{n}} \to N(0,1)$$

If we define $\overline{X} = \frac{1}{n} \sum_{i=1}^n X_i$, then

$$U_n = \frac{\sum_{i=1}^n X_i}{\sqrt{n}} = \frac{n\overline{X}}{\sqrt{n}} = \frac{\overline{X}}{1/\sqrt{n}} = \frac{\overline{X} - n\mu}{\sigma\sqrt{n}} \to N(0,1)$$

So we have $U_n \xrightarrow{L} N(0,1)$ by Central Limit Theorem.

Furthermore, if we define $Y_i = X_i^2$, then $V_n = \overline{Y}_n$, and we know that

$$\mathbb{E}(Y) = \mathbb{E}(X^2) = \mathbf{Var}(X) + (\mathbb{E}X)^2 = 1, \mathbf{Var}(Y) = \mathbb{E}(X^4) - (\mathbb{E}(X^2))^2 = \beta_4 - 1 < +\infty$$

And then we have

$$\mathbb{E}(\overline{Y}_n) = \mathbb{E}\left(\frac{Y_1 + \cdots + Y_n}{n} \right) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}Y_i = \mathbb{E}Y = 1$$

and

$$\mathbf{Var}(\overline{Y}_n) = \mathbf{Var}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2}\mathbf{Var}(Y_1 + \cdots + Y_n) = \frac{1}{n^2} n\,\mathbf{Var}(Y_i) = \frac{\mathbf{Var}(Y)}{n}.$$

So Chevbeshev's inequality states that for any $k > 0$,

$$\begin{aligned}\mathbb{P}\{|\overline{Y}_n - \mathbb{E}\overline{Y}_n| \geq k\} &\leq \frac{\mathbf{Var}(\overline{Y}_n)}{k^2} \\ &= \frac{\mathbf{Var}(Y)}{nk^2} \leq \frac{\beta_4}{nk^2} \to 0, \text{as } n \to +\infty.\end{aligned}$$

So we have that $\overline{Y}_n \xrightarrow{P} 1$, and by Cramer's theorem, we have

$$\frac{U_n}{V_n} \xrightarrow{L} N(0,1).$$

Example: Let $X_1, \cdots$ be a sequence of *i.i.d*, $U[0,1]$ random variables, and define

$$Y_n = \min\{X_1, \cdots, X_n\}$$

Then prove the following results separately:

(a) $Y_n \xrightarrow{d} 0$;

(b) $Y_n \xrightarrow{P} 0$;

(c) $Y_n \xrightarrow{a.s} 0$;

(d) $Y_n \xrightarrow{L^r} 0$ for all $r \geq 1$.

Chapter 6

References

[**1**]. *An Introduction to Probability and Statistics*, by Vijay K.Rohatgi and A.K. Md. Ehsanes Saleh.

[**2**]. *A First Course in Probability*, by Sheldon Ross.

[**3**]. *Introduction to Probability, Statistics and Random Processes*, by H. Pishro-Nik.

[**4**]. *Math 454 Course Notes, Fall 2024, McGill University*, by Prof. Linan Chen.

.

Thanks!