Introduction to Probability

Z.m. Wang[1]

1st Edition

[1] Research Institute of Economics and Management (RIEM), Southwest University of Finance and Economics, Chengdu, China. Email: zwang@swufe.edu.cn.

This manuscript is a preliminary edition. While efforts have been made to ensure its content is accurate, no warranty is made regarding its quality or correctness. Readers are encouraged to report any typographical errors or inaccuracies to the author for correction. The manuscript may also contain citations that have not been properly acknowledged, which will be addressed in future editions.

Preface

Embarking on the study of probability is both an opportunity and a challenge. This course represents a critical foundation for all subsequent studies in econometrics and data science. As we navigate an era of big data and artificial intelligence, proficiency in probability theory is indispensable. It not only builds the foundations for higher level courses but also cultivates a mindset geared towards probabilistic reasoning, which is essential for mitigating cognitive biases and making more informed decisions. Human cognition is often not well-equipped to handle probabilistic thinking. As we will see, human intuition and heuristics are mostly wrong about probabilistic events. This course seeks to provide a mathematical framework for properly understanding and applying probability.

Probability theory is more than just a mathematical discipline; it is a vital tool for making sense of uncertainty in the real world. Consider the myriad of questions that we encounter daily: Will it rain tomorrow? What is the expected return on an investment? What are the odds of a particular political party winning an election? How can a business optimize its customer service strategy when customer arrival times are unpredictable? Probability theory provides a scientific approach to answering these questions, enabling us to model and analyze uncertainties with mathematical tools.

However, this journey is not without its difficulties. For freshmen, particularly those new to calculus and linear algebra, the course presents a steep learning curve. The breadth of new concepts—such as random variables, expectations, and various distributions—can be overwhelming if encountered for the first time. Additionally, the use of advanced calculus, particularly integrals, may pose challenges for those who are still familiarizing themselves with these mathematical tools.

Despite these challenges, the rewards of studying probability are substantial.

Gaining a deep understanding of probability will not only enhance your knowledge base but also fundamentally transform your approach to problem-solving. The principles you will learn are applicable to a wide range of fields beyond econometrics and data science, including engineering, finance, and social sciences. You will learn the tools to approach these problems systematically and make informed decisions based on statistical evidence and probability. So be prepared for a challenging and rewarding journey!

Learning objectives

- **Review fundamental concepts:** Revisit the probability and calculus concepts learned in high school to ensure a solid foundation for more advanced topics.
- **Understand core probability theory:** Gain a thorough understanding of key concepts and theorems in probability theory, including random variables, expectations, covariances, and so on.
- **Develop probabilistic thinking:** Learn to approach problems with a probabilistic mindset and use random variables to describe and analyze uncertain outcomes.
- **Model real-world events:** Identify and apply important probability distributions to model and interpret real-world phenomena effectively.
- **Enjoy and have fun:** Discover and appreciate the inherent elegance of mathematics and the beauty of probability theory.

Study tips for new college students

- **Limit electronic distractions:** While digital tools like slides and tablets are convenient, traditional paper and pencil methods remain the most effective way to engage with and learn mathematics. Writing out problems and solutions helps reinforce concepts and improve retention.
- **Focus on key concepts:** College courses are often much more intensive than high school classes, and it is not feasible to master every detail. Concentrate on understanding the core ideas and principles, and don't get overwhelmed by the technical details.

- **Understand the “why”:** In mathematics, understanding the underlying reasons and logic behind methods is more important than just knowing how to do computations. The “why” helps you grasp the broader implications and applications of the techniques you learn.
- **Gain practical experience:** Although this course emphasizes theoretical understanding and does not require programming, experimenting with statistical software such as R can be highly beneficial. Practical experience with data manipulation and analysis will enhance your comprehension and stimulate interest in the subject.
- **Engage with the material:** I will strive to make the course engaging and less boring. However, if this course is not your primary interest, focus on the aspects of the material that intrigue you. Try to have a general impression of the major concepts even though you do not remember any detail.
- **Exams are important, but more important is to enjoy the course.**

This content of this book is organized or follows. We start with probabilities based on counting, which should be familiar to high school graduates. Though rudimentary, they often yield surprising results, as rigorous calculations frequently challenge our intuitions about the likelihood of events. Special emphasis is placed on conditional probability, as conditional thinking is crucial both in academic studies as well as in daily life.

Next, we introduce the core concept of the random variable, which forms the foundation of all probability distributions and statistical theory. Random variables are essential tools that allow us to mathematically model uncertainty. We introduce two types of random variables: discrete and continuous. We begin with discrete random variables because they do not require calculus, offering a smoother learning curve for beginners. Key concepts such as expectations, variance, and covariance are introduced alongside well-known discrete distributions such as the Binomial, Geometric, and Poisson distributions. This arrangement ensures that students can grasp these important concepts without being overwhelmed by calculus. We also demonstrate how these fundamental distributions can be applied to solve real-world problems.

Following this, we move on to continuous distributions. We will see that the formulas from discrete distributions extend naturally to continuous distributions with the aid of calculus—essentially replacing summation with integration. We cover some of the most important continuous distributions, such as the Normal, Exponential, and Gamma distributions, and explore the interconnections between them. We also extend the concepts of expectations, variance, and joint distributions to their continuous forms.

The book concludes with a discussion on sampling and statistical inference. Since we cannot observe entire distribution, it becomes necessary to infer distribution properties from finite samples. We introduce two of the most important theorems in probability and statistical theory—the Law of Large Numbers and the Central Limit Theorem. The breadth and generality of these theorems are remarkable. But their most significant contribution to statistical applications is they allow us to gauge how close our sample estimates are to the true parameter values. The final chapter also includes a brief discussion on estimator accuracy, confidence intervals, and hypothesis testing. These topics are introduced briefly, as they serve primarily to prepare students for more advanced courses, such as econometrics.

The chapters are organized logically, with each chapter building on the knowledge presented in the previous ones. Therefore, it is recommended to follow the sequence of chapters rather than reading them independently. However, advanced readers who are already familiar with the topics may feel free to skip between chapters as needed. This manuscript is written tersely, serving as a skeleton to complement lecture materials. It is not intended as a substitute for lectures or comprehensive textbooks. Students who wish to learn the course material solely by reading are encouraged to consult a formal textbook.

This manuscript is a preliminary version, and while efforts have been made to ensure accuracy, errors may still be present. Your feedback on any mistakes or inaccuracies is greatly appreciated and will help improve the material.

Contents

Chapter 1

Probability Basics

1.1 Introduction

Probability is the most important concept in modern science, especially as nobody has the slightest notion of what it means. ——Bertrand Russell

What is probability? We all talk about probabilities in everyday life, but mostly in vague languages. This course is to introduce probability as a logical framework for quantifying uncertainty and randomness.

Mathematics is the logic of certainty; probability is the logic of uncertainty.

The earliest development of probability is rooted in gambling. For instance, the renowned Monte Carlo method in statistics, invented by Stanislaw Ulam in the late 1940s, takes its name from the *Monte Carlo Casino* in Monaco, where Ulam's uncle would borrow money from relatives to gamble. Probability theories still apply today to analyze gambling odds, but their applications have expanded to nearly every field. It is the foundation of statistics, machine learning, and artificial intelligence. It also plays a crucial role in everyday decision-making, from stock investments to effective strategies to combat an infectious disease.

Probability is a concept that is intuitive to understand but very hard to define formally. Perhaps, the first formal definition of probability is often attributed to Pierre-Simon Laplace in the 18th century. In his work "Théorie analytique des probabilités," published in 1812,

> The probability of an event is the ratio of the number of cases favorable to it, to the number of all cases possible when nothing leads us to expect that any one of these cases should occur more than any other, which renders them, for us, equally possible.

This definition is outdated, as we will soon discover. But before we explore the modern definition of probability, let's first clarify some preliminary concepts (based on sets), which is the mathematical language we use to describe probabilistic events.

1.2 Events and sample spaces

The mathematical framework for probability is built around sets (like the cases in other math subjects as well).

Definition 1.1. The **sample space** S of an experiment is the set of all possible outcomes of the experiment. An **event** A is a subset of the sample space S. We say A occurred if the actual outcome is in A.

An experiment can be understood loosely. Anything (a gamble, an exam, a financial year, ...) can be an experiment. The sample space can be finite, countably infinite, or uncountably infinite. It is convenient to visualize events in a **Venn diagram**.

Set theory provides a rich language for expressing and working with events. Set operations, especially unions, intersections, and complements, make it easy to build new events in terms of already-defined events. For example, let S be the sample space of an experiment and let $A, B \subseteq S$ be events. Then the union $A \cup B$ is the event that occurs if and only if at least one of A and B occurs, the intersection $A \cap B$ is the event that occurs if and only if both A and B occur, and the complement A^c is the event that occurs if and only if A does not occur.

Example 1.1 (Coin flips)**.** A coin is flipped twice. We write "H" if a coin lands Head, and "T" if a coin lands Tail. The sample space is the set of all possible outcomes. Therefore, $S = \{HH, HT, TH, TT\}$. Let's look at some events:

1. Let A_1 be the event that the first flip is Heads. Then $A_1 = \{HH, HT\}$. Let A_2 be the event that the second flip is Heads. Then $A_2 = \{HH, TH\}$.

2. Let B be the event that at least one flip is Heads. Then $B = A_1 \cup A_2$.

3. Let C be the event that all the flips are Heads. Then $C = A_1 \cap A_2$.

4. Let D be the event that no flip is Heads. Then $D = B^c$.

Here is a list of events described in both English and set notations.

English	Sets
sample space	S
s is a possible outcome	$s \in S$
A is an event	$A \subseteq S$
A occurred	$s_{\text{actual}} \in A$
A or B	$A \cup B$
A and B	$A \cap B$
not A	A^c
at least one of $A_1, \ldots, A_n$	$A_1 \cup \cdots \cup A_n$
all of $A_1, \ldots, A_n$	$A_1 \cap \cdots \cap A_n$
A implies B	$A \subseteq B$
A and B are mutually exclusive (disjoint)	$A \cap B = \phi$
$A_1, \ldots, A_n$are a partition of S	$A_1 \cup \cdots \cup A_n = S$ and $A_i \cap A_j = \phi$ for $i \neq j$

1.3 Classical probability

Definition 1.2. Classical (naive) definition of probability:

$$P(A) = \frac{|A|}{|S|} = \frac{\text{number of outcomes favorable to A}}{\text{total number of outcomes in A}}$$

assuming the outcomes are *finite* and *equally likely.*

Example 1.2. Flip a coin twice. Find the probability of landing two heads.

Solution: There are four possible outcomes: {HH, HT, TH, TT}, each with equal probability. Therefore, $P(\text{HH}) = \frac{1}{4}$.

The naive definition is very restrictive. It has often been misapplied by people who assume equally likely outcomes without justification. Besides, it is easy to conceive examples of probabilities that do not fit into this formula, e.g. probability of rain. By saying it is "naive", it is definitely not the preferred definition in this course.

Nonetheless, we do some examples using the naive definition as a warm-up. Calculating the naive probability of an event A often involves counting the number of outcomes in A and the number of outcomes in the sample space S, which usually involve some counting methods. We now review some of the counting methods (multiplications, factorials, permutations, combinations) that was introduced in high schools.

Multiplications. Consider a compound experiment consisting of two sub-experiments, Experiment A and Experiment B. Suppose that Experiment A has a possible outcomes, and for each of those outcomes Experiment B has b possible outcomes. Then the compound experiment has $a \times b$ possible outcomes.

Exponentiations. Consider n objects and making k choices from them, one at a time with replacement. Then there are n^k possible outcomes.

Factorials. Consider n objects $1, 2, \ldots, n$. A permutation of $1, 2, \ldots, n$ is an arrangement of them in some order, e.g., $3, 5, 1, 2, 4$ is a permutation of $1, 2, 3, 4, 5$. The are $n!$ permutations of $1, 2, \ldots, n$.

Permutations. Consider n objects and making k choices from them, one at a time without replacement. Then there are $P_n^k = n(n-1)\cdots(n-k+1)$ possible outcomes, for $k \leq n$. (Ordering matters in this case, e.g. $1, 2, 3$ is considered different from $2, 3, 1$)

Combinations. Consider n objects and making k choices from them, one at a time without replacement, without distinguishing between the different orders in which they could be chosen (e.g. $1, 2, 3$ is considered no different from $2, 3, 1$). Then there are $C_n^k = \frac{n(n-1)\cdots(n-k+1)}{k!}$ possible outcomes. It literally counts the number of subsets of size k for a set of size n.

C_n^k is known as the Binomial coefficient, also denoted as $\binom{n}{k}$, read as "n choose k". As it is related to the Binomial theorem, which states that

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}.$$

The following table summarizes the counting methods.

	order matters	order doesn't matter
with replacement	n^k	C_{n+k-1}^k
non-replacement	P_n^k	C_n^k

We don't explain the upper-right corner case C_{n+k-1}^k as it is not relevant for our purpose here. Feel free to figure it out yourself if you are interested.

Example 1.3. Find the probability of a "full house" in a five-card hand.

Solution:

$$P(\text{Full House}) = \frac{13C_4^3 \cdot 12C_4^2}{C_{52}^5} = 0.14\%.$$

Example 1.4 (Birthday problem)**.** Suppose there are k people. Find the probability that two of them have the same birthday.

Solution: Assuming there are 365 days in a year, ignoring leap years. If $k > 365$, the probability is 1. If $k \leq 365$,

$$P(\text{no match}) = \frac{365 \cdot 365 \cdots (365 - k + 1)}{365^k};$$

$$P(\text{match}) = \begin{cases} 50.7\% & k = 23 \\ 70.6\% & k = 30 \\ 97\% & k = 50 \\ 99.999\% & k = 100 \end{cases}.$$

Example 1.5 (Newton-Pepys problem)**.** Isaac Newton was consulted about the following problem by Samuel Pepys, who wanted the information for gambling purposes. Which of the following events has the highest probability?

A: At least one 6 appears when 6 fair dice are rolled.

B: At least two 6's appear when 12 fair dice are rolled.

C: At least three 6's appear when 18 fair dice are rolled.

1.4 Axiomatic probability

We have now seen several methods for counting outcomes in a sample space, allowing us to calculate probabilities if the naive definition applies. But the naive definition can only take us so far, since it requires equally likely outcomes and can't handle an infinite sample space. To generalize the notion of probability, we'll use the best part about math, which is that you get to *make up your own definitions.* What this means is that we write down a short wish list of how we want probability to behave (in math, the items on the wish list are called axioms), and then we define a probability function to be something that satisfies the properties we want.

Definition 1.3. A **probability space** consists of S and P, where S is a sample space, and P is a function which takes an event $A \subseteq S$ as input and returns $P(A) \in [0, 1]$ such that

1. $P(\phi) = 0$,
2. $P(S) = 1$,
3. $P(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n)$ if $A_1, A_2, \ldots, A_n$ are disjoint.

Note that this Definition does not imply any particular interpretation of probability. In fact, any function P that satisfies the axioms are valid "probabilities". Thus, the theories of probability do not depend on any particular interpretation. It is purely axiomatic. From the three axioms, we can derive any property of probabilities. The interpretation also matters, but it is more of a philosophical debate. Basically, there are two views in this regard.

- The *frequentist* view of probability is that it represents a long-run frequency over a large number of repetitions of an experiment: if we say a coin has probability $1/2$ of Heads, that means the coin would land Heads 50% of the time if we tossed it over and over and over.
- The *Bayesian* view of probability is that it represents a degree of belief about the event in question, so we can assign probabilities to hypotheses like "candidate A will win the election" or "the defendant is guilty" even if it isn't possible to repeat the same election or the same crime over and over again.

Theorem 1.1. *Probability has the following properties. For any events A and B, we have*

1. $P(A^c) = 1 - P(A)$

2. If $A \subseteq B$, then $P(A) \leq P(B)$.
3. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Proof.

1. Since A and A^c are disjoint and their union is S, apply the third axiom:

$$P(S) = P(A \cup A^c) = P(A) + P(A^c);$$

By the second axiom, $P(S) = 1$. So $P(A) + P(A^c) = 1$.

2. The key is to break up the set into disjoint sets. If $A \subseteq B$, then $B = A \cup (B \cap A^c)$ where A and $B \cap A^c$ are disjoint (draw a Venn diagram for intuition). By the third axiom, we have

$$P(B) = P(A \cup (B \cap A^c)) = P(A) + P(B \cap A^c) \geq P(A).$$

3. We can write $A \cup B$ as the union of the disjoint set A and $B \cap A^c$. Then by the third axiom,

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

It suffices to show that $P(B \cap A^c) = P(B) - P(A \cap B)$. Since $B \cap A$ and $B \cap A^c$ are disjoint, we have

$$P(B) = P(B \cap A) + P(B \cap A^c).$$

So $P(B \cap A^c) = P(B) - P(A \cap B)$ as desired. □

The last property is a very usueful formula for finding the probability of a union of events when the events are not necessarily disjoint. We have showed that for two events A and B. A natural question is to generalize it for three or more events. For three events,

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

We skip the proof. It can be easily justified by showing a Venn diagram. For the n-events case, we state it as the following theorem.

Theorem 1.2 (Inclusion-exclusion)**.** *For any events* $A_1, A_2, \ldots, A_n$*, it holds that*

$$P(A_1 \cup A_2 \cdots \cup A_n) = \sum_{j=1}^{n} P(A_j) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) - \cdots$$
$$(-1)^{n+1} P(A_1 \cap \cdots \cap A_n).$$

This formula can be proved by induction using the axioms. Below is a famous application (known as de Montmort's problem, named after French mathematician Pierre Remond de Montmort) of the inclulsion-exclusion theorem.

Example 1.6 (Matching problem)**.** Given n cards, labeled $1, 2, ..., n$. Let A_j be the event "j-th card matches"(the j-th card is numbered as j). Find the probability of at least one match, i.e. $P(A_1 \cup A_2 \cup \cdots \cup A_n) =?$

Solution: Since all position are equally likely, $P(A_j) = \frac{1}{n}$. The probability of there being two matches is: $P(A_1 \cap A_2) = \frac{(n-2)!}{n!} = \frac{1}{n(n-1)}$. Similarly, the probability of there being k matches is: $P(A_1 \cap \cdots \cap A_k) = \frac{(n-k)!}{n!} = \frac{1}{n(n-1)\cdots(n-k+1)}$. Using the property of the union of events,

$$\begin{aligned} P(A_1 \cup A_2 \cup \cdots \cup A_n) =& n \cdot \frac{1}{n} - \binom{n}{2} \frac{1}{n(n-1)} + \binom{n}{3} \frac{1}{n(n-1)(n-2)} - \cdots \\ =& 1 - \frac{1}{2!} + \frac{1}{3!} - \frac{1}{4!} + \cdots + (-1)^{n+1} \frac{1}{n!} \approx 1 - \frac{1}{e}. \end{aligned}$$

1.5 Conditional probability

Abraham Wald, the renowned statistician, was hired by the Statistical Research Group (SRG) at Columbia University to figure out how to minimize the damage to bomber aircraft. The data they had comprised aircraft returning from missions with bullet holes on their bodies. If asked which parts of the aircraft should be armored to enhance survivability, the obvious answer seemed to be to armor the damaged parts. However, Wald suggested the exact opposite—to armor the parts that were not damaged. Why? Because the observed damage was conditioned on the aircraft returning. If an aircraft had been damaged on other parts, it likely would not have returned. Thinking conditionally completely changes the answer![1]

The probability of A **conditioned on** B is the updated probability of event A after we learn that event B has occurred. Since events contain information, the occurring of a certain event may change our believes on probabilities of other relevant events. The updated probability of event A after we learn that event B has occurred is the conditional probability of A given B.

Definition 1.4. If A and B are events with $P(B) > 0$, then the **conditional probability** of A given B is defined as

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

Exercise 1.1. Prove that conditional probabilities are probabilities. (Hint: using the three axioms.)

Theorem 1.3. *Properties of conditional probability:*

- $P(A \cap B) = P(B)P(A|B) = P(A)P(B|A)$
- $P(A_1 \cap \cdots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2) \cdots P(A_n|A_1 \ldots A_{n-1})$
- $P(A|B) = \dfrac{P(B|A)P(A)}{P(B)}$ *(Bayes' rule)*

The last property, Bayes' rule, quantifies how to update probabilities based on new evidence. It is named after Thomas Bayes in the 18th century. It gained prominence posthumously through Richard Price's publication of Bayes' work in 1763. The rule calculates the probability of a hypothesis based on prior knowledge and new data, foundational for Bayesian statistics.

Historically, Bayes studied the problem in order to prove David Hume wrong. Hume argued that we cannot directly observe causation; instead, we infer it from patterns of events. Bayes' rule allows for a systematic way to update our beliefs about causal relationships as new evidence emerges, thereby bridging the gap between empirical observation and theoretical inference. This approach counters Hume's skepticism by providing a method for rationally assessing the likelihood of causes based on observed effects.[2]

[1] See an interesting talk by Professor Joseph Blitzstein: "The Soul of Statistics". Available on `http://www.youtube.com/watch?v=dzFf3r1yph8`

[2] See `https://faculty.som.yale.edu/jameschoi/bayes-theorem-began-as-a-defense-of-christianity`.

Theorem 1.4 (Law of total probability)**.** *Let $A_1, ..., A_n$ be a partition of the sample space S (i.e., the A_i are disjoint events and their union is S), with $P(A_i) > 0$ for all i. Then*

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i).$$

Example 1.7. Get a random 2-card hand from a standard deck. Find the probability of getting another ace conditioned on (a) having one ace, or (b) having the ace of spade.

Solution: The example shows the subtleness of conditional probabilities. The seemingly indifferent probabilities are in fact different:

$$P(\text{another ace} \mid \text{one ace}) = \frac{P(\text{both aces})}{P(\text{one ace})} = \frac{C_4^2/C_{52}^2}{1 - C_{48}^2/C_{52}^2} = \frac{1}{33};$$
$$P(\text{another ace} \mid \text{ace of spade}) = \frac{P(\text{ace of spade \& another ace})}{P(\text{ace of spade})} = \frac{C_3^1/C_{52}^2}{C_{51}^1/C_{52}^2} = \frac{1}{17}.$$

In the first case, the denominator is interpreted as "at least one ace"; whereas in the second case, it is "ace of space + another card".

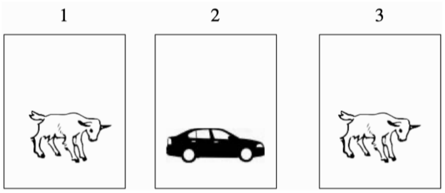
Example 1.8. The pandemic afflicted roughly 1/3 of the world population. The PCR test is 98% accurate. (this means if you have been infected, the test reports positive 98% of the time.) Find the probability of being infected when a test is positive.

Solution: Let D: actually infected, T: test positive. The test accuracy means: $P(T|D) = 98\%$. It also means $P(T|D^C) = 2\%$. We also know that $P(D) = 1/3$. We want to find $P(D|T)$. Apply the Bayes' rule:

$$\begin{aligned} P(D|T) =& \frac{P(T|D)P(D)}{P(T)} \\ =& \frac{P(T|D)P(D)}{P(T|D)P(D) + P(T|D^C)P(D^C)} \\ =& \frac{0.98 \times 1/3}{0.98 \times 1/3 + 0.02 \times 2/3} \approx 96\%. \end{aligned}$$

Note that how $P(T|D)$ is different from $P(D|T)$, though confusing the conditionality is quite common in daily life. The difference is even pronounced if the disease is rare. Suppose $P(D) = 10\%$. Then $P(D|T) = 84\%$. A large difference from the test accuracy rate 98%!

Example 1.9 (Monty Hall problem)**.** Suppose you are on Monty Hall's TV show. There are three doors. One of them has a car behind it. The other two doors have goats. Monty knows which one has the car. Monty now asks you to pick one door. You will win whatever is behind the door. After you pick one door. Monty opens another door that shows a goat. Monty then asks you if you want to switch. Is it optimal to switch?



We present two solutions to the problem. The first one is using the law of total probability. Let S: succeed assuming switch; D_j: door j has the car, $j \in 1, 2, 3$. Without loss of generality, assume the initial pick is Door 1. Monty will always open the door with a goat. By the law of total probability,

$$\begin{aligned} P(S) = & \underbrace{P(S|D_1)}_{\text{switch from initial pick}} P(D_1) + \underbrace{P(S|D_2)}_{\text{Monty opens door 3}} P(D_2) + \underbrace{P(S|D_3)}_{\text{Monty opens door 2}} P(D_3) \\ =& 0 + 1 \times \frac{1}{3} + 1 \times \frac{1}{3} = \frac{2}{3}. \end{aligned}$$

The problem can also be solved using the Bayes' rule. Let D_j: door j has the car; M_j: Monty opens door j, $j \in 1, 2, 3$. Assume the initial pick is Door 1. If Monty opens door 3, the probability of winning the car assuming switching is

$$\begin{aligned} P(D_2|M_3) =& \frac{P(M_3|D_2)P(D_2)}{P(M_3)} \\ =& \frac{P(M_3|D_2)P(D_2)}{P(M_3|D_1)P(D_1) + P(M_3|D_2)P(D_2) + P(M_3|D_3)P(D_3)} \\ =& \frac{1 \times \frac{1}{3}}{\frac{1}{2} \times \frac{1}{3} + 1 \times \frac{1}{3} + 0} = \frac{2}{3}. \end{aligned}$$

Note that, if door 1 has the car, Monty will open door 2 and 3 with equal probability, thus $P(M_3|D_1) = \frac{1}{2}$. And Monty will never open the door with the car, therefore $P(M_3|D_3) = 0$. Similarly, if Monty opens door 2, we have $P(D_3|M_2) = \frac{2}{3}$. Therefore, the optimal choice is always to switch. Intuitively, because Monty knows which door has the car, the fact that he always opens the door without the car gives additional information regarding the choice of the door.

Example 1.10 (Simpson's paradox)**.** There are two doctors, Dr. Lee and Dr. Wong, performing two types of surgeries — heart surgery (hard) and band-aid removal (easy). Dr. Lee has higher overall surgery success rate. Is Dr. Lee necessarily a better doctor than Dr. Wong?

Solution: No. Consider the following example:

	Dr. Lee			Dr. Wong		
	Heart	Band-Aid	Total	Heart	Band-Aid	Total
Success	2	81	83	70	10	80
Failure	8	9	17	20	0	20
Success rate	20%	90%	83%	78%	100%	80%

The truth is Dr. Lee has overall higher success rate because he only does easy surgeries (band-aid removal). Dr. Wong does mostly hard surgeries and thus has a lower overall success rate. Yet, he is better at each single type of surgery. To formalize the argument, let S: successful surgery; D: treated by Dr. Lee, D^c: treated by Dr. Wong; E: heart surgery, E^c: band-aid removal. Dr. Wong is better at each type of surgery,

$$\begin{aligned} P(S|D,E) &< P(S|D^c,E) \\ P(S|D,E^c) &< P(S|D^c,E^c); \end{aligned}$$

But, Dr. Lee has a higher overall successful rate,

$$P(S|D) > P(S|D^c).$$

This is because there is a "confounder" E:

$$P(S|D) = \underbrace{P(S|D,E)}_{<P(S|D^c,E)} \underbrace{P(E|D)}_{\text{weight}} + \underbrace{P(S|D,E^c)}_{<P(S|D^c,E^c)} \underbrace{P(E^c|D)}_{\text{weight}}.$$

A **confounder** is a variable that influences with both explanatory variable and the outcome variable, which therefore "confounds" the correlation between the two. In our example, the type of surgery (E) is associated with both the doctor and the outcome. Without the confounder being controlled, it is impossible to draw valid conclusions from the statistics.

In general terms, Simpson's paradox refers to the paradox in which a trend that appears across different groups of aggregate data is the reverse of the trend that appears when the aggregate data is broken up into its components. It is one of the most common sources of statistical misuse. Here is another example.[3]

Example 1.11 (UC Berkeley gender bias)**.** One of the best-known examples of Simpson's paradox comes from a study of gender bias among graduate school admissions to University of California, Berkeley. The admission figures for the fall of 1973 showed that men applying were more likely than women to be admitted, and the difference was so large that it was unlikely to be due to chance.

	Male		Female	
	Applicants	Admitted	Applicants	Admitted
Total	8,442	44%	4,321	35%

However, when taking into account the information about departments being applied to, the conclusion turns to the opposite: in most departments, the admission rate for women is higher than men. The lower overall admission rate is caused by the fact that women tended to apply to more competitive departments with lower rates of admission, whereas men tended to apply to less competitive departments with higher rates of admission.

Department	Male		Female	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%
Total	2691	45%	1835	30%

[3] See https://setosa.io/simpsons for a really good illustration of the Simpson's paradox.

1.6 Independence

Definition 1.5. If event B's occurrence does not change the probability of A, then we say A and B are independent. That is to say A and B are **independent** if

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

Or more generally, A and B are **independent** if

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

(A definition including cases where A or B has zero probability.)

Theorem 1.5. *If events A and B are independent, then*

- *A and B^c are independent;*
- *A^c and B^c are independent.*

A and B are independent means they do not provide information to each other in the sense that conditional probability is not different from the unconditional probability. It is not an intuitive idea as it seems. It will become clearer when we discuss random variables in later chapters. Here we clarify some likely confusions.

Remark 1.1. Independence is not the same as disjointness.

A and B are disjoint means if A occurs, B cannot occur. But independence means A occurs has nothing to do with B.

Remark 1.2. Pairwise independence does not imply independence.

Definition 1.6. Events A, B, and C are said to be **(mutually) independent** if all of the following equations hold:

$$\begin{aligned} P(A \cap B) =& P(A)P(B), \\ P(A \cap C) =& P(A)P(C), \\ P(B \cap C) =& P(B)P(C), \\ P(A \cap B \cap C) =& P(A)P(B)P(C). \end{aligned}$$

If the first three conditions hold, we say that A, B, and C are **pairwise independent**. Pairwise independence does not imply independence. Convince yourself with the following example.

Example 1.12. Consider two fair, independent coin tosses, and let A be the event that the first is Heads, B the event that the second is Heads, and C the event that both tosses have the same result. Show that A, B, and C are pairwise independent but not independent.

Solution: For each event, $P(A) = \frac{1}{2}$, $P(B) = \frac{1}{2}$. Consider the two events together, there are four possible outcomes: HH, HT, TH, TT. $P(C) = P(HH) + P(TT) = \frac{1}{2}$. Thus,

$$\begin{aligned} P(A \cap B) = P(HH) &= \frac{1}{4} = P(A)P(B) \\ P(A \cap C) = P(HH) &= \frac{1}{4} = P(A)P(C) \\ P(B \cap C) = P(HH) &= \frac{1}{4} = P(B)P(C) \end{aligned}$$

But A, B, C are not independent, because

$$P(A \cap B \cap C) = P(HH) = \frac{1}{4} \neq P(A)P(B)P(C).$$

Definition 1.7. For n events $A_1, A_2, \ldots, A_n$ to be **(mutually) independent**, we require any pair to satisfy $P(A_i \cap A_j) = P(A_i)P(A_j)$ (for $i \neq j$), any triplet to satisfy $P(A_i \cap A_j \cap A_k) = P(A_i)P(A_j)P(A_k)$ (for i, j, k distinct), and similarly for all quadruplets, quintuplets, and so on.

Definition 1.8. Events A and B are **conditional independent** given C if

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

Remark 1.3. Conditional independence does not apply independence.

Consider an example of playing chess games. Conditioned on the strength of your opponents, the outcome of each game is reasonably independent (ignoring the psychology and fatigues of the players). But the outcomes are not unconditionally independent, because stronger player has higher chances of winning each game.

Remark 1.4. Independence does not apply conditional independence.

Consider an example of fire alarm. Suppose there are two potential causes to trigger the fire alarm: (1) there is fire; (2) someone smoking. Assume the two events are independent. But they are not conditional independent if conditioning on the alarm beeping. Because if the alarm is on, but no one smokes, we definitely know there is fire. So there they are not conditional independent.

1.7 Review of calculus*

Calculus is a prerequisite to work with continuous distributions. The following chapters assume readers are proficient in calculus. We nonetheless review some basic concepts here as a warm-up. This review is not exhaustive, so please refer to a specific textbook if needed for a more comprehensive understanding.

1.7.1 Differentiation

We define the derivative of a function $f(x)$ to be

$$f'(x) = \lim_{h\to 0} \frac{f(x+h) - f(x)}{h}$$

Loosely speaking, a function is continuous if there is no jump in the graph, differentiable if the curve is smooth. Some commonly used derivatives:

$$\begin{aligned}
\frac{d}{dx}(x^n) &= nx^{n-1} \\
\frac{d}{dx}(e^x) &= e^x \\
\frac{d}{dx}(\ln(x)) &= \frac{1}{x} \\
\frac{d}{dx}(\sin(x)) &= \cos(x) \\
\frac{d}{dx}(\cos(x)) &= -\sin(x) \\
(fg)' &= f'g + fg' \\
\left(\frac{f}{g}\right)' &= \frac{f'g - fg'}{g^2} \\
[f(g(x))]' &= f'(g(x))g'(x)
\end{aligned}$$

When dealing with limits of the form "$\frac{0}{0}$" or "$\frac{\infty}{\infty}$", the L'Hospital rule is very handy.

$$\lim_{x \to a} \frac{f(x)}{g(x)} = \lim_{x \to a} \frac{f'(x)}{g'(x)}.$$

One important application of derivatives is the Taylor's theorem, which gives the approximation of a function around a given point by polynomials. Assume function f is at least k times differentiable, then

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(k)}(a)}{k!}(x-a)^k + \cdots$$

which means we can approximate a function arbitrarily well by higher order polynomials. Some commonly used Taylor series (expanding around $a = 0$):

$$\begin{aligned}
\frac{1}{1-x} &= 1 + x + x^2 + x^3 + \cdots \quad \text{for } |x| < 1 \\
e^x &= 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots \\
\sin x &= x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots \\
\cos x &= 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots \\
\ln(1+x) &= x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots \quad \text{for } |x| < 1 \\
\arctan(x) &= x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \cdots \quad \text{for } |x| \leq 1
\end{aligned}$$

Taylor series are one of the most amazing results in calculus. For example, in the last formula, if we let $x = 1$:

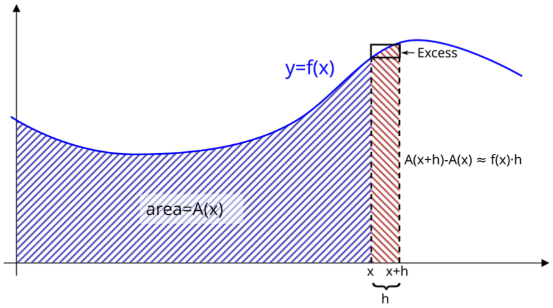
$$\frac{\pi}{4} = \arctan(1) = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots$$

Therefore, we can approximate π by summing up a sequence of fractions:

$$\pi = 4\left(1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots\right).$$

1.7.2 Integration

Integration is the inverse operation of differentiation. Integral has the geometric interpretation as the area under the curve. Let $A(x)$ be the area under the curve of $y = f(x)$. Thus $A(x) = \int_0^x f(t)dt$. The change of the area resulted from a tiny little change of x is approximated by $A(x+h) - A(x) \approx f(x)h$. That is $\frac{A(x+h)-A(x)}{h} = f(x)$. If the change is infinitesimal, $h \to 0$, we have $A'(x) = f(x)$.



The Fundamental Theorem of Calculus: if F is the antiderivative of f, then

$$F(x) = \int_a^x f(t)dt$$

$$\int_a^b f(x)dx = F(b) - F(a)$$

One interpretation of the integral is — the integral of a rate of change of a quantity gives the net change in that quantity. Think about speed and distance: $\int_a^b v(t)dt = s(b) - s(a)$.

Because the integral is just a sum over infinitely many approximating rectangles, $\int_a^b f(x)dx = \lim_{n\to\infty} \sum_{i=1}^n f(x_i)\Delta x$. Integrals behave just like sums. For example, $\frac{1}{b-a}\int_a^b f(x)dx$ has the interpretation of the average of $f(x)$ from a to b.

Indefinite integrals are the general antiderivatives without specifying the interval of the integration. It always comes with a constant C. Some commonly used integrals:

$$
\begin{aligned}
\int dx &= x + C \\
\int x^n dx &= \frac{x^{n+1}}{n+1} + C \\
\int e^x dx &= e^x + C \\
\int \frac{1}{x} dx &= \ln|x| + C \\
\int \cos(x) dx &= \sin(x) + C \\
\int \sin(x) dx &= -\cos(x) + C \\
\int \frac{1}{1+x^2} dx &= \arctan(x) + C
\end{aligned}
$$

Two common integration techniques are *substitution* and *integration by parts*. Integration by substitution applies when substituting part of the integrand makes the integral more approachable.

Example 1.13. Find $\int \sqrt{3x+2}dx$.

Solution: Let $u = 3x + 2$, then $du = 3dx$. Then

$$\int \sqrt{3x+2}dx = \frac{1}{3}\int \sqrt{u}du = \frac{2}{9}u^{3/2} + C = \frac{2}{9}(3x+2)^{3/2} + C.$$

Integration by parts follows the formula: $\int f(x)g'(x)dx = f(x)g(x) - \int f'(x)g(x)dx$.

Example 1.14. Find $\int x \sin x dx$.

Solution: Let $f(x) = x$, $g'(x) = \sin x$. Then $g(x) = -\cos x$. Therefore,

$$\int x \sin x dx = -x \cos x - \int (-\cos x)dx = -x\cos x + \sin x + C.$$

Chapter 2

Random Variables

2.1 Introduction

In the previous chapter, we have been working with *events*, which is a conceptualization of real world outcomes occurred with probabilities. In this chapter, we introduce a much more powerful conceptualization that deals with uncertain outcomes — random variables, which is the foundation of all probability and statistical studies.

In high school, all mathematical models come with certainty. For example, the falling time of any object from height h down to the earth is: $t = \sqrt{\frac{2h}{g}}$, where g is the gravity constant. The outcome is *deterministic.* The variables that enter into the equation either have unknown values or known certain values. Errors are possible only due to frictions or measurement errors.

But many real world processes come naturally with uncertainty. Think about the temperature tomorrow, or the stock market returns. We can only make predictions with probabilities. Yes, you may argue this uncertainly is due to incomplete information. If we have all the knowledge regarding the climate, we can predict exactly the temperature. But given the imperfection of the human knowledge, the only feasible option is to build this uncertainly into our mathematical models. Random variable is core concept and the Swiss knife that we use to deal with uncertainties mathematically.

Informally, a random variable differs from a normal variable as it is "random".

> A random variable is a variable whose value is uncertain, but comes with probabilities.

A random variable, say X, is never associated with a certain value, such as $X = 1$, or $X = 2$. It could be any of these values, but with different probabilities, e.g. $P(X = 1) = 0.2$, $P(X = 2) = 0.4$.

Definition 2.1. Given an experiment with sample space S, a **random variable** is a function from the sample space S to the real numbers $\mathbb{R}$.

As an example, flipping a coin twice, let X be the number of heads. Then $X(\cdot)$ is a functions that maps events in $\{HH, HT, TH, TT\}$ into real numbers. In our case, the mapping goes like

$$X(HH) = 2, X(HT) = 1, X(TH) = 1, X(TT) = 0.$$

X is therefore an <u>encoding</u> of events in the sample space into real numbers. We could, of course, have different encodings. Conder the random variable Y as the number of tails. Then we have $Y = 2 - X$.

$$Y(HH) = 0, Y(HT) = 1, Y(TH) = 2, Y(TT) = 2.$$

We could also define Z as the number heads in the 1st toss only. The encoding goes like

$$Z(HH) = 1, Z(HT) = 1, Z(TH) = 0, Z(TT) = 0.$$

We have listed three ways of "encoding" the same experiment as random variables. All of them are valid random variables, but they map the outcomes into different numbers. We can say that, a random variable is a <u>numeric</u> "summary" of an aspect of an experiment.

Remark. We usually use capital letters, such as X, Y, Z, to denote random variables. We use small letters, such as x, y, z, to denote specific values. $P(X = x)$ means the probability of X taking the value x. "$X = x$" is an event. In the example above, $X = 2$ corresponds to the event HH. Note that we don't write $P(X)$. It is meaningless if X takes no value.

Definition 2.2. Let X be a random variable. The **distribution** of X is the collection of all probabilities of the form $P(X \in C)$ for all sets C of real numbers such that $\{X \in C\}$ is an event.

A **distribution** specifies the probabilities associated with all values of a random variable. In the above example, the distribution of X is given by

$$P(X=0)=\frac{1}{4}, P(X=1)=\frac{1}{2}, P(X=2)=\frac{1}{4}.$$

The distribution of Y is given by

$$P(Y=0)=\frac{1}{4}, P(Y=1)=\frac{1}{2}, P(Y=2)=\frac{1}{4}.$$

The distribution of Z is given by

$$P(Z=0)=\frac{1}{2}, P(Z=1)=\frac{1}{2}.$$

You may have noted that the probabilities in a distribution always sums up to 1, as all possible events constitute the entire sample space.

Example 2.1. Roll two fair 6-sided dice. Let $T = X + Y$ be the total of the two rolls, where X and Y are the individual rolls. Find the distribution for T.

2.2 Discrete random variables

Definition 2.3. We say X is a **discrete random variable** if X can take only a finite number k of different values $x_1, \ldots, x_k$ or, at most, an infinite sequence of countable different values $x_1, x_2, \ldots$.

The finite or countably infinite set of values x such that $P(X = x) > 0$ is called the **support** of X.

Definition 2.4. If a random variable X has a discrete distribution, the **probability mass function** (PMF, sometimes also known as **probability function**, or **frequency function**) of X is defined as the function p such that

$$p(x) = P(X = x)$$

where $p(x) \geq 0$ for all possible values of x and $\sum_{\text{all } x} p(x) = 1$.

Remark. $p(x)$ differs from the probability function $P(\cdot)$. $p(x)$ is a real-valued function. We can manipulate it as normal real-valued functions. Some textbooks prefer to use $f(x)$. In this book, we use $p(x)$ to distinguish it from the probability density function for continuous random variables. Sometimes, it is convinient to add a subscript, $p_X(x)$, to specify this is the PMF for random variable X.

Remark. The PMF $p(x)$ of a random variable X must satisfy the following criteria:

- Nonnegative: $p(x) \geq 0$ for all possible values of x;
- Sums to 1: $\sum_{\text{all } x} p(x) = 1$.

There are different ways to represent a PMF. We can (1) list all the possible values and their associated probabilities; (2) write a formula for the PMF; or (3) visualize it in a graph.

2.3 Continuous random variables

Definition 2.5. We say a random variable X has a **continuous distribution** if the possible values of X takes the form of a continuum.

Definition 2.6. For a continuous random variable X, the **probability density function** (PDF) of X is a real-valued function f such that

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

where $f(x) \geq 0$ for all x and $\int_{-\infty}^{+\infty} f(x)dx = 1$.

Continuous random variables are usually measurements. Examples include height, weight, temperature, the amount of money and so on.

Remark. PDF differs from the discrete PMF in important ways:

- For a continuous random variable, $P(X = x) = 0$ for all x;
- The quantity $f(x)$ is not a probability. To get the probability, we integrate the PDF (probability is the area under the PDF):

$$P(a < X \leq b) = F(b) - F(a) = \int_a^b f(x)dx.$$

- Since any single value has probability 0, including or excluding endpoints does not matter.

$$P(a < X < b) = P(a < X \leq b) = P(a \leq X < b) = P(a \leq X \leq b).$$

Remark. The PDF f of a continuous random variable must satisfy the following criteria:

- Nonnegative: $f(x) \geq 0$;
- Integrates to 1: $\int_{-\infty}^{\infty} f(x)dx = 1$.

More on continuous distributions will be discussed in later chapters.

2.4 Cumulative distribution function

Definition 2.7. The **cumulative distribution function** (CDF) of a random variable X is the function F given by $F(x) = P(X \leq x)$.

For discrete random variables, $F(x) = \sum_{k \leq x} p(k)$.

For continuous random variables, $F(x) = \int_{-\infty}^{x} f(t)dt$. We thus have $\frac{dF(x)}{dx} = f(x)$.

Unlike PMF or PDF, a cumulative distribution function can be defined for both discrete and continuous random variables. CDF gives the full distribution of a random variable. Given the CDF, we can figure out any probability distribution of the random variable. For example, $P(x_1 < x \leq x_2) = F(x_2) - F(x_1)$.

Theorem 2.1. Any CDF has the following properties:

- $P(X > x) = 1 - F(x)$
- $P(x_1 < x \leq x_2) = F(x_2) - F(x_1)$
- Increasing: if $x_1 \leq x_2$, then $F(x_1) \leq F(x_2)$.
- Right-continuous: for any a, $F(a) = \lim_{x \to a+} F(x)$.
- $F(x) \to 0$ as $x \to -\infty$; $F(x) \to 1$ as $x \to +\infty$.

2.5 Data and random variables

Imagine you're standing in a bustling café, sipping your coffee while observing the scene around you. Customers come and go, some ordering their usual drinks, others trying something new. The number of people who walk in during an hour, the time each spends waiting in line, even the total sales for the day—these are all examples of things we can measure, and all are influenced by uncertainty. We understand these seemingly unpredictable happenings through random variables and their distributions.

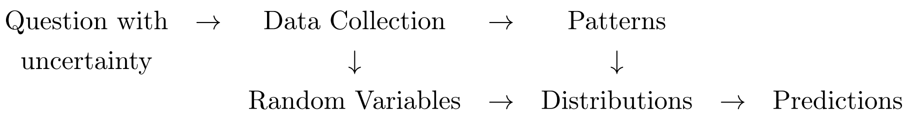
A random variable is a mathematical abstraction that provides a bridge between theoretical probability and real-world data. Every dataset you encounter—whether it's student grades, daily temperatures, or stock market prices—can be viewed as observations from random variables.

The fact is, we describe something as a random variable not because they are random in nature (like rolling a die), but it is the problem posed does not allow a definite answer (the question itself involves likelihood) or we do not have enough information to give an answer with certainty. In such cases, we would like to give the possible values with their chances of being (the idea of distribution).

Despite the outcome of any one event being uncertain, we can use patterns from past observations to predict the general behavior of these variables. By collecting data, we can figure out how often certain outcomes happen and connect them to theoretical models called **distributions**.

For instance, if you track the heights of 100 people, you might notice that most are close to the average, with fewer at the extremes. This "bell-shaped curve" is the hallmark of something called the **normal distribution**, one of the most common patterns in nature. Heights, test scores, and even measurement errors tend to follow this distribution.

But not all data fits the same shape. If you're counting the number of cars passing through an intersection in a given hour, you might find the **Poisson distribution** is a better fit. This pattern shows up whenever you're dealing with counts of events over time or space—like customer arrivals at a store or typos in a book.



Question with uncertainty → Data Collection → Patterns

Data Collection ↓ Random Variables → Distributions → Predictions

Patterns ↓ Distributions

By linking real-world data to these theoretical models, random variables let us make predictions. How many customers will show up next week? What's the chance of a traffic jam during rush hour? Random variables give us the tools to answer these questions with confidence.

Viewing the world through the lens of random variables has several benefits: (1) It helps us deal with uncertainty. Random variables give us a framework to understand situations where outcomes aren't guaranteed. For example, a meteorologist uses random variables to estimate the chance of rain, so you know whether to carry an umbrella. (2) It connects theory to reality. By analyzing data, we can identify which theoretical models describe the randomness we observe. This helps businesses, scientists, and policymakers make better decisions. (3) It allows for better planning.

Suppose you're running an online store. Knowing that the number of daily orders follows a certain distribution can help you manage inventory and staffing. Suppose you're tracking the number of customers visiting your café each day. You notice the number fluctuates between 50 and 100, with an average of 75. By treating this as a random variable, you can estimate the likelihood of having fewer than 60 customers tomorrow (useful for planning staff schedules); or the probability of exceeding 90 customers on a holiday (important for stocking supplies).

In the grand scheme of things, random variables are more than just mathematical tools—they're a way to make sense of life's unpredictability. So, the next time you see data—whether it's sports stats, sales figures, or even your social media likes—remember: Behind the numbers is a pattern of randomness waiting to be understood. And with random variables, you have the key to unlock it.

2.6 Summary

1. A random variable serves as a numerical representation of a specific aspect of an experiment or a random phenomenon. It allows us to quantify outcomes

in a meaningful way, enabling analysis and interpretation of the results. For example, in a coin toss, we might define a random variable to represent the number of heads observed in a series of flips.

2. We typically model situations as random variables because we often lack sufficient information to draw definitive conclusions. In these instances, probability provides a framework for making educated guesses about uncertain outcomes. It acts as a compromise, allowing us to express our uncertainty mathematically and make decisions based on incomplete information. This is particularly useful in fields like finance, economics, and social sciences, where uncertainties are inherent.

3. Generally, we do not have access to the true distribution of a random variable, which is why we rely on finite samples, often derived from historical records, to approximate this distribution. By analyzing past data, we can estimate the probabilities associated with different outcomes. However, it's important to note that these approximations are subject to sampling variability and may not capture the entire complexity of the underlying phenomenon. Thus, understanding the limitations of our data and the potential for biases is crucial when making inferences based on random variables.

Chapter 3

Discrete Distributions

3.1 Bernoulli distribution

We introduce some theoretical distributions from now on. These distributions are important because they provide standardized models for common "patterns" of random processes. We develop these distributions from idealized assumptions. In practice, we usually do not know what is the "true" distribution of the question of interest. Typically, we "fit" the question into a theoretical distribution according to their proximity to the assumptions.

Definition 3.1. A random variable X is said to have the **Bernoulli distribution**, denoted as $X\sim$ Bern(p), if X has only two possible values, 0 and 1, and $P(X=1)=p$, $P(X=0)=1-p$.

The PMF of a Bernoulli random variable X is given by

$$p_X(x)=\begin{cases} p & \text{if } x=1, \\ 1-p & \text{if } x=0. \end{cases}$$

This can also be expressed as

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

Example 3.1. The Bernoulli distribution is widely used because it provides a simple yet powerful framework for modeling binary outcomes, where events can be classified as *success* or *failure* (**Bernoulli trial**). This versatility allows it to be applied across a wide range of fields and scenarios.

One key reason for its popularity is that many real-world phenomena can be distilled into binary outcomes. For instance, in quality control, a product can either pass or fail inspection; in healthcare, a treatment may either be effective or ineffective; and in marketing, a consumer may either purchase a product or not. Because nearly any situation involving two possible outcomes can be framed in terms of success and failure, the Bernoulli distribution becomes a natural choice for analysis.

Example 3.2. An **indicator variable** assigns a value of 1 to represent the occurrence of a specific event (success) and a value of 0 to indicate that the event did not happen (failure). This binary representation allows us to convert any event into a random variable,

$$I_A = \begin{cases} 1 & A \text{ happens with probability } p \\ 0 & \text{otherwise, with probability } 1-p \end{cases}$$

which can then be modeled as a Berboulli distribution $I_A \sim Bern(p)$.

Bernoulli distribution serves as the foundation for more complex models, such as the binomial distribution, which deals with multiple independent trials. This hierarchical structure makes it easier to build upon and develop more sophisticated statistical methods. Its simplicity also facilitates calculations and interpretations, making it accessible for researchers and practitioners alike.

3.2 Binomial distribution

Definition 3.2. Suppose $X_1, X_2, \ldots, X_n$ are independent and identical Bern(p) distributions. Let X be the total number of successes of the n independent trials. That is, $X = X_1 + X_2 + \cdots + X_n$. Then X has the **Binomial distribution**, $X \sim \text{Bin}(n, p)$.

The probability mass function of X directly follows from the combination theory:

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

This is a valid PMF because, by the Binomial theorem, we have

$$\sum_{k=0}^{n} p_X(k) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = (p + (1-p))^n = 1.$$

Example 3.3. In the previous example of tossing two coins, we compute the distribution of X by counting the equally likely outcomes in an event. We can get the same result by realizing it is a Binomial distribution. $X \sim \mathrm{Bin}(2, 1/2)$. Since each coin tossing is an independent Bernoulli trial. The probabilities come directly from the PMF.

$$\begin{aligned}
P(X=0) = p_X(0) &= \binom{2}{0}\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^2 = \frac{1}{4};\\
P(X=1) = p_X(1) &= \binom{2}{1}\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^1 = \frac{1}{2};\\
P(X=2) = p_X(2) &= \binom{2}{2}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^0 = \frac{1}{4}.
\end{aligned}$$

Utilizing the Binomial distribution also allows us to generalize the problem. Suppose we are tossing n coins, we want to find the probability of getting k heads. It is almost impossible to count all the possible outcomes, but the answer immediately follows from the Binomial PMF.

Example 3.4. The Binomial distribution is often used to model the probability that a certain number of "successes" occur during a certain number of trials. Here is an example. Suppose it is known that 5% of adults who take a certain medication experience negative side effects. We want to find the probability that a certain number of patients in a random sample of 100 will experience negative side effects. Let X be the number patients that experience negative side effects, it follows that $X \sim \mathrm{Bin}(100, 0.05)$.

Example 3.5. Let $X \sim Bin(n, p)$ and $Y \sim Bin(m, p)$ be two independent Binomail random variables. Show that $X + Y \sim Bin(n + m, p)$.

Proof. By the definition of the Binomial distribution, X is the number of successes in n independent trials, and Y is the number of successes in m independent trials. Therefore, $X + Y$ is the number of successes in $n + m$ independent trials, which is exactly $Bin(n + m, p)$.

We can also prove it using indicator variables. $X = \sum_{i=1}^{n} X_i$ where $X_i \sim Bern(p)$; $Y = \sum_{j=1}^{m} Y_j$ where $Y_j \sim Bern(p)$. Therefore, $X + Y = \sum_{i=1}^{n} X_i +$

$\sum_{j=1}^{m} Y_j = \sum_{k=1}^{n+m} Z_k$. Since X_i and Y_j are identical Bernoulli random variables, $Z_k = X_k$ for $k = 1, \ldots, n$; $Z_k = Y_{k-n}$ for $k = n+1, \ldots, n+m$.

Another way is to leverage the PMF:

$$\begin{aligned}
P(X+Y=k) &= \sum_{i+j=k} P(X=i)P(Y=j) \\
&= \sum_{i+j=k} \binom{n}{i} p^i (1-p)^{n-i} \binom{m}{j} p^j (1-p)^{m-j} \\
&= \sum_{i+j=k} \binom{n}{i}\binom{m}{j} p^{i+j} (1-p)^{m+n-i-j} \\
&= p^k (1-p)^{m+n-k} \sum_{i=0}^{k} \binom{n}{i}\binom{m}{k-i} \\
&= p^k (1-p)^{m+n-k} \binom{n+m}{k}.
\end{aligned}$$

The last step: $\binom{n+m}{k} = \sum_{i=0}^{k} \binom{n}{i}\binom{m}{k-i}$ is known as the Vandermonde's identity.

Example 3.6. Let's explore an example that appears to be Binomial but is, in fact, not a Binomial distribution. Given a 5-card hand. Find the distribution of the number of aces.

Solution. Let X be the number of aces. It is tempting to say $X \sim Bin(5, p)$. But this not correct. Because having one ace is NOT independent from having another ace. We need to use the classical approach:

$$P(X=k) = \frac{C_4^k C_{48}^{5-k}}{C_{52}^5}.$$

This example leads to a named distribution that is closed related to Binomial — Hypergeometric distribution.

3.3 Hypergeometric distribution

Suppose we have a box filled with w white and b black balls. We draw n balls out of the box with replacement. Let X be the number of white balls. Then $X \sim Bin(n, w/(w+b))$. Since the draws are independent Bernoulli trials, each with probability $w/(w+b)$ of success. If we instead sample without replacement, then the number of white balls follows a **Hypergeometric distribution**. We denote this by $X \sim \text{HGeom}(w, b, n)$.

Theorem 3.1. *If $X \sim HGeom(w, b, n)$, then the PMF of X is*

$$p_X(k) = \frac{\binom{w}{k}\binom{b}{n-k}}{\binom{w+b}{n}},$$

for integers k satisfying $0 \leq k \leq w$ and $0 \leq n-k \leq b$, and $p_X(k) = 0$ otherwise.

Example 3.7. Let's redo the ace-card exercise with Hypergeometric distribution. In a 5-card hand, the number of aces in the hand has the $\text{HGeom}(4, 48, 5)$ distribution, which can be seen by thinking of the aces as white balls and the non-aces as black balls. The probability of having exactly three aces is $\frac{\binom{4}{4}\binom{48}{2}}{\binom{52}{5}} = 0.0017\%$.

The Binomial and Hypergeometric distributions are often confused. Both are discrete distributions taking on integer values between 0 and n for some n, and both can be interpreted as the number of successes in n Bernoulli trials. However, a crucial part of the Binomial story is that the Bernoulli trials involved are independent. The Bernoulli trials in the Hypergeometric story are dependent, since the sampling is done without replacement.

3.4 Geometric and Negative Binomial

Definition 3.3. Consider a sequence of independent Bernoulli trials, each with the same success probability p. Let X be the number of failures before the first successful trial. Then X has a **Geometric distribution**, $X \sim \text{Geom}(p)$.

Let's derive the PMF for the Geometric distribution. By definition,

$$P(X = k) = q^k p$$

where $q = 1 - p$. This is a valid PMF because

$$\sum_{k=0}^{\infty} q^k p = p \sum_{k=0}^{\infty} q^k = \frac{p}{1-q} = 1.$$

The expectation of X is given by

$$E(X) = \sum_{k=0}^{\infty} k \cdot q^k p = p \sum_{k=0}^{\infty} kq^k = p\frac{q}{p^2} = \frac{q}{p}.$$

To see why this holds, taking derivative with respect to q on both sides of $\sum_{k=0}^{\infty} q^k = \frac{1}{1-q}$ yields

$$\sum_{k=1}^{\infty} kq^{k-1} = \frac{1}{(1-q)^2};$$

Then multiply both sides by q:

$$\sum_{k=1}^{\infty} kq^k = \frac{q}{(1-q)^2} = \frac{q}{p^2}.$$

A generalization of the Geometric distribution is the Negative Binomial distribution.

Definition 3.4. In a sequence of independent Bernoulli trials with success probability p, if X is the number of failures before the r-th success, then X is said to have a **Negative Binomial distribution**, denoted $X \sim \text{NBin}(r, p)$.

The PMF for Negative Binomial distribution, by definition, is given by

$$P(X = k) = \binom{k+r-1}{r-1} q^k p^r.$$

To compute the expectation, let $X = X_1 + \cdots + X_r$ where X_i is the number of failures between the $(i-1)$-th success and the i-th success, $1 \leq i \leq r$. Then $X_i \sim \text{Geom}(p)$. By linearity of expectations,

$$E(X) = E(X_1) + \cdots + E(X_r) = r\frac{1-p}{p}.$$

Example 3.8 (Toy collector)**.** There are n types of toys. Assume each time you buy a toy, it is equally likely to be any of the n types. What is the expected number of toys you need to buy until you have a complete set?

Solution: Define the following random variables:

$$\begin{aligned}
T =& T_1 + T_2 + \cdots + T_n \\
T_1 =& \text{number of toys until 1st new type} \\
T_2 =& \text{additional number of toys until 2nd new type} \\
T_3 =& \text{additional number of toys until 3rd new type} \\
&\vdots
\end{aligned}$$

We know, $T_1 = 1$, $T_2 - 1 \sim$ Geom $\left(\frac{n-1}{n}\right)$,..., $T_j - 1 \sim$ Geom $\left(\frac{n-(j-1)}{n}\right)$. Thus,

$$\begin{aligned}
E(T) =& E(T_1) + E(T_2) + \cdots + E(T_n) \\
=& 1 + \frac{n}{n-1} + \frac{n}{n-2} + \cdots + \frac{1}{n} \\
=& n(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}) \\
\rightarrow& n(\log n + 0.577).
\end{aligned}$$

If $n = 5$, $E(T) \approx 11$; if $n = 10$, $E(T) \approx 29$.

3.5 Discrete uniform distribution

Definition 3.5. Let $a \leq b$ be integers. Suppose that the value of a random variable X is equally likely to be each of the integers $a, \ldots, b$. Then we say that X has the **discrete uniform distribution** on the integers $a, \ldots, b$. We denote it as $X \sim DUnif(a, \ldots, b)$.

The PMF of $X \sim DUnif(a, \ldots, b)$ is given by

$$p(x) = \begin{cases} \frac{1}{b-a+1} & \text{for } x = a, \ldots, b \\ 0 & \text{otherwise} \end{cases}.$$

Example 3.9. Let X be a random number from 1,2,...,100. Then $X \sim DUnif(1, ..., 100)$. And $P(X = k) = 1/100$ for any $k = 1, ..., 100$.

The uniform distribution can be defined in discrete cases, but its continuous form is more well-known.

Definition 3.6. Let a and b be two real numbers such that $a < b$. Let X be a random variable such that $a \leq X \leq b$ and, for every subinterval interval of $[a, b]$, the probability that X belongs to that subinterval is proportional to the length of that subinterval. Then we say X has the **uniform distribution** on the interval $[a, b]$. We denote it as $X \sim Unif(a, b)$.

The PDF of $X \sim Unif(a, b)$ is given by

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

We verify this is a valid PDF because

$$\int_a^b f(x)dx = \int_a^b \frac{1}{b-a}dx = \frac{1}{b-a}\int_a^b dx = 1;$$

or the area of the rectangle surrounded by $x = a, x = b$ and $f(x) = \frac{1}{b-a}$ is 1.

3.6 Functions of random variables

Functions of random variables are also random variables. If X is a random variable, then X^2, e^X and $\sin(X)$ are also random variables.

Definition 3.7. For an experiment with sample space S, a random variable X, and a function $g : \mathbb{R} \to \mathbb{R}$. $g(X)$ is the random variable that maps s to $g(X(s))$ for all $s \in S$.

Theorem 3.2. *Let X be a discrete random variable and $g : \mathbb{R} \to \mathbb{R}$. If $g(X)$ is a one-to-one function. Then the support of $g(X)$ is the set of all y such that $x = g^{-1}(y)$ is in the support of X. The PMF of $g(X)$ is*

$$P(g(X) = y) = P(g(X) = g(x)) = P(X = x).$$

Theorem 3.3. *Let X be a discrete random variable and $g : \mathbb{R} \to \mathbb{R}$. Then the support of $g(X)$ is the set of all y such that $g(x) = y$ for at least one x in the support of X. The PMF of $g(X)$ is*

$$P(g(X) = y) = \sum_{x:g(x)=y} P(X = x).$$

Definition 3.8. Give an experiment with sample space S, if X, Y are random variables that map $s \in S$ to $X(s)$ and $Y(s)$, then $g(X, Y)$ is the random variable that maps s to $g(X(s), Y(s))$ for all $s \in S$.

Example 3.10. We roll two fair 6-sided dice. Let X be the number on the first die, and Y be the number on the second die. Find the distribution of $\max(X, Y)$.

Solution: We just show how to compute one case in the distribution, other cases are similar.

$$\begin{aligned} P(\max(X,Y) = 5) &= P(X = 5, Y \le 4) + P(X \le 4, Y = 5) + P(X = 5, Y = 5) \\ &= 2P(X = 5, Y \le 4) + 1/36 \\ &= 2(4/36) + 1/36 = 9/36. \end{aligned}$$

3.7 Independence of random variables

Definition 3.9. Random variables X and Y are **independent** if

$$P(X \le x, Y \le y) = P(X \le x)P(Y \le y)$$

for all $x, y \in \mathbb{R}$.

In the discrete case, this is equivalent to the condition

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

for all possible values of x, y.

In the continuous case, this is equivalent to the condition

$$f(x, y) = f_X(x) f_Y(y)$$

for all x, y, where f, f_X, f_Y are density functions.

Theorem 3.4. *If X and Y are independent, then any function of X is independent of any function of Y.*

Definition 3.10. If a given number of random variables are independent and have the same distribution, we call them **independent and identically distributed**, or **i.i.d** for short.

- Independent and identically distributed (X, Y independent die rolls)
- Independent and not identically distributed (X: die roll; Y: coin flip)
- Dependent and identically distributed (X: number of Heads; Y: number of Tails)
- Dependent and not identically distributed (X: economic growth; Y: presidential election)

Definition 3.11. Random variables X and Y are **conditionally independent** given Z if

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

for all $x, y \in \mathbb{R}$ and all z in the support of Z.

For discrete random variables, the equivalent definition is to require

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

3.8 Application: seller ratings*

This example involves multiple types of discrete distributions. The technique used to solve this problem aligns with Bayesian inference, which is beyond the scope of this course. However, it remains an interesting case. The procedure illustrates the process of statistical modeling: we begin with an assumption and a proposed statistical model, then update it with new data. Finally, we draw inferences based on the model, typically addressing the question we aim to answer. You are not required to understand everything in this example. Nonetheless, it helps to develop a mindset of statistical inference early in the study.

Suppose you are shopping a product online. There are three sellers with the following ratings:

- Seller 1: 100% positive out of 10 reviews
- Seller 2: 96% positive out of 50 reviews
- Seller 3: 93% positive out of 200 reviews

Which seller is likely to give the best service?

The problem is intriguing because it is obvious that higher ratings do not necessarily means higher satisfaction. We have to weight in the number of reviews. The more reviews, the more trustworthy the ratings are. Let $X_j^{(i)}$ be a random variable that means consumer j is satisfied with seller i, where $i \in \{1, 2, 3\}$. Assume $X_j^{(i)}$ follows a Bernoulli distribution:

$$X_j^{(i)} = \begin{cases} 1 & \text{satisfaction with probability } \theta_i \\ 0 & \text{otherwise} \end{cases}$$

where θ_i is an unknown parameter of seller i that captures their "genuine" satisfaction rate. We assume the consumers independently write their ratings. The overall positive rate of seller i is therefore $R_i = \frac{1}{n_i} \sum_j X_j^{(i)}$ where n_i is the total number of reviews. We want to infer the value of θ_i from their observed positive rate R_i. From now on we drop the seller index i to simply the notation since it is symmetric for all sellers.

Because we have no prior knowledge about θ. We assume that θ takes any value from $[0, 1]$ equally likely, i.e. $\theta \sim \text{Unif}(0, 1)$. Assuming each X_j is independent and identical, then

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

follows the Binomial distribution with PMF:

$$p(k|\theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k}$$

Our goal is to find: $p(\theta|k)$. Recall that the Bayes' rule allows us to invert the conditional probability:

$$p(\theta|k) = \frac{p(k|\theta)p(\theta)}{p(k)} = \frac{p(k|\theta)p(\theta)}{\int_{-\infty}^{\infty} p(k|\theta)p(\theta)d\theta}$$

Since $\theta \sim \text{Unif}(0,1)$, we have

$$p(\theta) = \begin{cases} 1 & \text{if } \theta \in [0,1] \\ 0 & \text{otherwise} \end{cases}$$

We now focus on $\theta \in [0,1]$, since the probability is 0 otherwise. Substitute in the PMF of the Binomial distribution,

$$p(\theta|k) = \frac{\binom{n}{k}\theta^k(1-\theta)^{n-k}}{\int_0^1 \binom{n}{k}\theta^k(1-\theta)^{n-k}d\theta}$$

The hard part is to evaluate the integral. We state without proof (this is known as the Beta function, which we will prove in later chapters):

$$\int_0^1 \theta^k(1-\theta)^{n-k} = \frac{k!(n-k)!}{(n+1)!}$$

Therefore,

$$p(\theta|k) = \frac{(n+1)!}{k!(n-k)!}\theta^k(1-\theta)^{n-k}$$

Now suppose you are the next customer. The probability that you would be satisfied is

$$\begin{aligned} P(X_{n+1} = 1|S = k) &= \int_0^1 P(x_{n+1} = 1|\theta)p(\theta|k)d\theta \\ &= \int_0^1 \theta \times \frac{(n+1)!}{k!(n-k)!}\theta^k(1-\theta)^{n-k}d\theta \\ &= \frac{(n+1)!}{k!(n-k)!}\int_0^1 \theta^{k+1}(1-\theta)^{(n+1)-(k+1)}d\theta \\ &= \frac{(n+1)!}{k!(n-k)!} \times \frac{(k+1)!(n-k)!}{(n+2)!} \\ &= \frac{k+1}{n+2}. \end{aligned}$$

Now we substitute the ratings for the three sellers:

- Seller 1: $n = 10, k = 10$

- Seller 2: $n = 50, k = 48$
- Seller 3: $n = 200, k = 186$

The probabilities that you would be satisfied with each seller are: 92%, 94%, 93%. The result is known as the **Laplace's rule of succession**. The rule of thumb is, pretending we have too more reviews: one is positive, the other is negative. Compute the satisfaction rate as $\frac{k+1}{n+2}$.

3.9 Further comments

How to choose a distribution?

The fact is, we never know the "true" distribution of a real-world problem. When building a probability model, the distribution is typically *assumed* based on the nature of the data and the problem at hand. This assumption is crucial because the probability distribution determines how the random variable behaves, including its likelihood of taking specific values. Typically, this process involves:

1. Choosing the distribution: Based on the characteristics of the real-world situation or data, you assume an appropriate probability distribution. For example, if you're modeling the number of successes in a fixed number of independent trials, you might assume a Binomial distribution.
2. Assumptions behind the distribution: Every distribution has underlying assumptions. For example, a Binomial distribution assumes independent trials with two possible outcomes (success/failure) and a constant probability of success.
3. Fitting the model: Once you assume a distribution, you use data to estimate parameters of the distribution (e.g., mean, variance, or rate parameters), which allows you to make probabilistic predictions and inferences.

It is important to stress that the data we have collected from real events does not directly reveal the **Data Generating Process (DGP)**, which is the true underlying process that produces the data. Instead, when we assume a distribution, we are essentially making a hypothesis about what that DGP might be.

The actual relationship between the assumed distribution and the data is one of approximation and testing, rather than perfect correspondence.

The assumed distribution is a *theoretical model* that we believe could explain the underlying patterns in the data. The data is a finite set of observations, which is only a sample from the potential infinite population or DGP. The data is influenced by noise, randomness, and sample size, so it doesn't always clearly show the true DGP. When we assume a distribution, we're making an educated guess about the DGP based on the nature of the problem, properties of the data, and sometimes prior knowledge or experience.

Data alone, especially from a finite sample, does not directly tell us what the DGP is. Instead, we infer the DGP by fitting models to the data and assessing how well they describe it. Since data is inherently noisy and finite, different models may fit the data well, meaning that multiple distributions could seem plausible based on the data alone. That's why we use goodness-of-fit tests, residual analysis, and model comparison to narrow down our choices.

If the data pattern conflicts with the assumed distribution, it might suggest that the assumption be wrong, and we should revisit our model. However, some degree of mismatch can be due to sample noise, outliers, or oversimplification, and may not always mean the assumption is entirely incorrect.

The workflow of probability modeling

The example of sellers' ratings is a good illustration of how we do probability modeling. Here we summarize it into several key steps.

1. Understanding of the problem and data exploration: The typical workflow of probability modeling begins with a clear understanding of the problem we are trying to solve. This involves identifying the objective of the model, determining which quantities or events need to be modeled as random variables. This also involves gathering relevant data, if available, or understanding the kind of data we will be working with.

2. Assumption of probability distribution: Based on the nature of the data and the problem at hand, choose a candidate distribution. For discrete data, this could be distributions like Bernoulli or Binomial. For continuous data, it might be Normal or Uniform distributions.

3. Parameter estimation: The candidate distribution usually involves unknown parameters. In most of the applications, we are interested in estimating these parameters. In our example, we update the parameter with the Bayes' rule. But there are other estimation methods available, such as Maximum Likelihood Estimation (MLE). Estimation quantifies the model and provides specific estimates based on the data.

4. Model fit and evaluation: We skip this step in our example. But normally, we need to evaluate how well the assumed distribution fits the data. This involves performing goodness-of-fit tests or graphical diagnostics. If the assumed distribution doesn't fit the data well, the model might need to be refined.

5. Simulation or inference: After refining the model, we can run simulations or make inferences. If the model is meant to simulate real-world processes, we can now generate new data based on the probability distribution and its parameters. We may also use the model to predict future outcomes or estimate probabilities of specific events.

Chapter 4

Expectation

4.1 Expectation

Definition 4.1. Let X be a discrete random variable. The **expectation** of X, denoted by $E(X)$, is defined as:

$$E(X) = \sum_{\text{all } x} xP(X = x).$$

The expectation of X is also referred to as the **mean** of X or the **expected value** of X.

In other words, the expected value of X is a *weighted average* of the possible values that X can take on, weighted by their probabilities. If the values are of equal probability, expectation is the simple average of all x: $E(X) = \frac{1}{n} \sum x$.

The expected value of X is a *number* (if it exists), $E(X) \in \mathbb{R}$. It is not a random variable such as a function of X.

Sometimes, we would like to omit the parentheses for simplicity and write $EX \equiv E(X)$. We also like to denote expectation by the greek letter $\mu \equiv E(X)$.

Example 4.1. The expectation of a Bernoulli random variable $X \sim \text{Bern}(p)$:

$$E(X) = 1 \times P(X = 1) + 0 \times P(X = 0) = p.$$

Example 4.2. The expectation of a Binomial random variable $X \sim \text{Bin}(n, p)$:

$$\begin{aligned}
E(X) =& \sum_{k=0}^{n} kp(k) \\
=& \sum_{k=0}^{n} k \cdot \binom{n}{k} p^k q^{n-k} \\
=& \sum_{k=1}^{n} n \cdot \binom{n-1}{k-1} p^k q^{n-k} \\
=& np \sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} q^{n-k} \\
=& np \underbrace{\sum_{j=0}^{n-1} \binom{n-1}{j} p^j q^{n-1-j}}_{\text{another Binomial PMF}} \\
=& np.
\end{aligned}$$

Example 4.3. Life expectancy is the average number of years a person is expected to live. It is a crucial indicator of the quality of living and one of the three components of the Human Development Index (HDI) (the other two components are education and per capita GDP). Here is a toy example to compute life expectancy with hypothetical data.[1]

(1)	(2)	(3)	(4)		(5)	(6)
Age	Population	Mortality rates	# Survive		# Died at age	P(Age)
0	200	1%	1000		10	1%
20	300	2%	990	=1000(1-1%)	20	2%
40	250	10%	970	=990(1-2%)	97	10%
60	150	20%	873	=970(1-10%)	175	17%
80	100	100%	699	=873(1-20%)	699	70%
Total	1000					

Table 4.1: Hypothetical mortality rates and life table

To simplify our analysis, we will assume there are only five possible ages: 0, 20, 40, 60, and 80. A baby is born at age 0, and can either die at that age or

[1] This is an overly simplified example that only serves to clarify the definition of expectation. See this tutorial from MEASURE Evaluation for the actual computation of life expectancy.

survive to age 20. We intentionally exclude intermediate ages such as 5 and 10 for the sake of computational simplicity.

It's important to note that life expectancy is not the same as the average age of the population. For instance, based on the hypothetical data presented, the average age can be calculated as:

$$\overline{\text{Age}} = (0 \times 200 + 20 \times 300 + 40 \times 250 + 60 \times 150 + 80 \times 100)/1000 = 33.$$

However, the expected age, denoted as $E(\text{Age})$, is defined as:

$$E(\text{Age}) = \sum \text{Age} \times P(\text{Age}).$$

To compute this expected value, we need to determine $P(\text{Age})$, the probability of living to a specific age or dying at that age. This requires consideration of the mortality rate at each age, which is given in Column 3.

Assuming 1000 babies are born at age 0, with a mortality rate of 1% at that age, we find that 99% of the babies survive to age 20. Thus, the number of babies that survive to age 20 is: $1000 \times (1 - 1\%) = 990$. We can apply similar calculations to determine the number of survivors at each subsequent age.

The number of individuals who die at a specific age (Column 5) is the difference between the number of survivors at that age and the next (Column 4). To find the probability of living to a specific age, we compute: $P(\text{Age}) =$ Column 4/1000.

Finally, we compute the expected value of age (or life expectancy) as follows:

$$E(Age) = 0 \times 1\% + 20 \times 2\% + 40 \times 10\% + 60 \times 17\% + 80 \times 70\% = 70.6.$$

This figure differs from the average age. Since the mortality rate is low at younger ages, the probabilities $P(\text{Age})$ for these ages are also low, while they are higher for older ages. This example illustrates the distinction between average and expected values. In everyday conversation, we may use these terms interchangeably, but in certain contexts, expected values can significantly differ from averages.

4.2 Linearity of expectation

Theorem 4.1. *For any random variables X, Y and any constant c,*

$$\begin{aligned} E(X+Y) &= E(X)+E(Y), \\ E(cX) &= cE(X). \end{aligned}$$

This property holds regardless of the dependencies between the random variables.

Proof. The proof is not as straightforward as it seems. It is hard to combine the two random variables:

$$E(X)+E(Y) = \sum_x xP(X=x)+\sum_y yP(Y=y) \overset{?}{=} \sum (x+y)P(X+Y=x+y).$$

The problem becomes easier if the number of possible values of X and Y are the same and all values are equally likely,

$$E(X)+E(Y) = \frac{1}{n}\sum x + \frac{1}{n}\sum y = \frac{1}{n}\sum (x+y) = E(X+Y).$$

The original problem is equivalent to the simple case if realizing that the weighted average is jut a simple average with repetitive values. For example,

$$1 \times \frac{1}{4} + 2 \times \frac{2}{4} + 3 \times \frac{1}{4} = \frac{1}{4}(1+2+2+3).$$

Imagine the sample space as being composed of "atom" outcomes $\{\omega\}$, each with equal probability $P(\omega)$. All random variable are function of these atoms, $X(\omega)$, and $Y(\omega)$. Therefore, the expectation formula can be rewritten as

$$E(X)+E(Y) = \sum_\omega X(\omega)P(\omega)+\sum_\omega Y(\omega)P(\omega) = \sum_\omega (X+Y)(\omega)P(\omega) = E(X+Y).$$

Here is another way to prove linearity for discrete random variables:

$$\begin{aligned}
E(X+Y) &= \sum_{z=x+y} zP(X+Y=z) \\
E(X+Y) &= \sum_x \sum_y (x+y)P(X=x,Y=y) \\
&= \sum_x \sum_y xP(X=x,Y=y) + \sum_x \sum_y yP(X=x,Y=y) \\
&= \sum_x x \sum_y P(X=x,Y=y) + \sum_y y \sum_x P(X=x,Y=y) \\
&= \sum_x xP((X=x) \cap \bigcup_{\text{all } y} (Y=y)) + \sum_y yP(\bigcup_{\text{all } x} (X=x) \cap (Y=y)) \\
&= \sum_x xP(X=x) + \sum_y yP(Y=y) \\
&= E(X) + E(Y).
\end{aligned}$$

□

Corollary 4.1. *Further properties on the linearity of expectations:*

- If $Y = aX + b$, then $E(Y) = aE(X) + b$.
- $E(X_1 + \cdots + X_n) = E(X_1) + \cdots + E(X_n)$
- $E(a_1X_1 + \cdots + a_nX_n + b) = a_1E(X_1) + \cdots + a_nE(X_n) + b$

Example 4.4. Redo the expectation of $X \sim \text{Bin}(n,p)$ with properties of expectation:

$$E(X) = E(X_1 + \cdots + X_n) = nE(X_i) = np$$

where $X_i \sim \text{Bern}(p)$.

Example 4.5. Let $X \sim \text{HGeom}(w,b,n)$. Find $E(X)$ the expected number of white balls. Similarly, we can decompose X:

$$X = I_1 + \cdots + I_n$$

where I_j equals 1 if the jth ball is white and 0 otherwise. We have said that $\{I_j\}$ are not independent, but the property of linearity still holds:

$$E(X) = E(I_1 + \cdots + I_n) = E(I_1) + \cdots + E(I_n).$$

Meanwhile we have

$$E(I_j) = P(j\text{-th ball is white}) = \frac{w}{w+b}$$

since unconditionally the jth ball is equally likely to be any of the balls. Thus, $E(X) = \frac{nw}{w+b}$.

Example 4.6. In a group of n people, what is the expected number of distinct birthdays among the n people (the expected number of days on which at least one of the people was born)? What is the expected number of people sharing a birthday (any day)?

Solution: Let X be the number of distinct birthdays, and write $X = I_1 + \cdots + I_{365}$, where

$$I_j = \begin{cases} 1 & \text{if someone was born on day } j \\ 0 & \text{otherwise} \end{cases}.$$

Then

$$\begin{aligned} E(I_j) &= P(\text{someone was born on day } j) \\ &= 1 - P(\text{no one was born on day } j) \\ &= 1 - \left(\frac{364}{365}\right)^n. \end{aligned}$$

Then by linearity,

$$E(X) = 365\left(1 - \left(\frac{364}{365}\right)^n\right).$$

Let Y be the number of people sharing a birthday, and $Y = J_1 + \cdots + J_n$ where J_k is an indicator that the j-th person shares his birthday with somebody else.

$$\begin{aligned} E(J_k) &= P(\text{someone shares birthday with } k) \\ &= 1 - P(\text{no one shares birthday with } k) \\ &= 1 - \left(\frac{364}{365}\right)^{n-1}. \end{aligned}$$

Therefore,

$$E(Y) = \sum_{k=1}^{n} E(J_k) = n\left(1 - \left(\frac{364}{365}\right)^{n-1}\right).$$

For some numeric values, $E(Y) = 2.3$ if $n = 30$; $E(Y) = 6.3$ if $n = 50$.

Example 4.7. Suppose that there are n people sitting in a classroom with exactly n seats. At some point, everyone got up, ran around the room, and sat back down randomly (i.e., all seating arrangements are equally likely). What is the expected value of the number of people sitting in their original seat?

Solution: Number the people from 1 to n. Let X_i be the Bernoulli random variable with value 1 if person i returns to their original seat and value 0 otherwise. Since person i is equally likely to sit back down in any of the n seats, the probability that person i returns to their original seat is $1/n$. Therefore $E[X_i] = 1/n$. Now, let X be the number of people sitting in their original seat following the rearrangement. Then $X = X_1 + X_2 + \cdots + X_n$. By linearity of expected values, we have $E[X] = \sum E[X_i] = \sum 1/n = 1$.

Example 4.8. Let Π be a permutation over $\{1, 2, \ldots, n\}$. That is a reordering of the numbers. A fixed point of a permutation are the points not moved by the permutation. For example, in the permutation below

$$\begin{array}{ccccc} & 1 & 2 & 3 & 4 \\ \Pi & 2 & 4 & 3 & 1 \end{array}$$

The fixed point is 3. Find the expected number of fixed points of a random permutation.

Solution: Let X be the number of fixed points of a random permutation. Then $X = \sum_{k=1}^{n} \mathbf{1}_{\Pi(k)=k}$ where $\mathbf{1}_{\Pi(k)=k}$ indicates the k-th number stays the same after the permutation. By linearity,

$$E(X) = E\left(\sum_{k=1}^{n} \mathbf{1}_{\Pi(k)=k}\right) = \sum_{k=1}^{n} E\left(\mathbf{1}_{\Pi(k)=k}\right) = \sum_{k=1}^{n} \frac{1}{n} = 1.$$

Example 4.9 (Buffon's needle)**.** Rule a surface with parallel lines a distance d apart. What is the probability that a randomly dropped needle of length $l \leq d$ crosses a line?

Solution: Consider dropping any (continuous) curve of length l onto the surface. Imagine dividing up the curve into N straight line segments, each of length $\frac{l}{N}$. Let X_i be the indicator for the i-th segment crossing a line. Let X be the total number of times the curve crosses a line. Then,

$$E(X) = E(\sum X_i) = \sum E(X_i) = N \cdot E(X_i).$$

There could be infinitely many segments. It is hard to compute this expectation directly. But here we arrive an important Lemma: the expected number of crossings is proportional to the length of the curve, regardless of the shape of the curve. If we can compute $E(X)$ for some curve, the we can compute $E(X)$ for any length by scaling the value proportional to the length.

Consider a circle of diameter d. The circle always crosses the lines twice for sure. That is, $E(X_{\text{circle}}) = 2$. The length of the circle is πd. Therefore, the value of $E(X)$ for any curve of length l is given by

$$E(X) = \frac{2l}{\pi d}.$$

Now a needle can cross a line either 1 or 0 times. Thus, $E(X) = 1 \cdot P(X = 1) + 0 \cdot P(X = 0)$ is exactly the probability of a needle crossing a line.

Remark. This amazing example can be used to approximate the value of π. Let q be the probability of a needle crossing a line. q can be approximated by large number of simulations. Then $\pi \approx \frac{2l}{qd}$.

4.3 Multiplication and LOTUS

Theorem 4.2. If X and Y are independent, we have

$$E(XY) = E(X)E(Y).$$

In general, if $X_1, \dots, X_n$ are independent, we have

$$E(X_1 X_2 \cdots X_n) = E(X_1)E(X_2) \cdots E(X_n).$$

Remark. The multiplication rule will not hold without independence.

Proof. For discrete and independent X, Y,

$$\begin{aligned} E(XY) &= \sum_x \sum_y xyP(X = x, Y = y) \\ &= \sum_x \sum_y xyP(X = x)P(Y = y) \quad \text{if independent} \\ &= \sum_x xP(X = x) \sum_y yP(Y = y) \\ &= E(X)E(Y). \end{aligned}$$

□

Remark. This is a sufficient but not necessary condition. $E(XY) = E(X)E(Y)$ does not imply independence. Consider a counter-example,

$$X = \begin{cases} 1 & \text{with prob. } 1/2 \\ 0 & \text{with prob. } 1/2 \end{cases}, \quad Z = \begin{cases} 1 & \text{with prob. } 1/2 \\ -1 & \text{with prob. } 1/2 \end{cases};$$

Then

$$Y = XZ = \begin{cases} -1 & \text{with prob. } 1/4 \\ 0 & \text{with prob. } 1/2 \\ 1 & \text{with prob. } 1/4 \end{cases}.$$

We have $E(X) = 1/2$, $E(Y) = 0$, $E(XY) = 0$. So $E(XY) = E(X)E(Y)$. But clearly X, Y are not independent.

Theorem 4.3 (Law of the unconscious statistician (LOTUS))**.** *Let X be a random variable, and g be a real-valued function of a real variable. If X has a discrete distribution, then*

$$E[g(X)] = \sum_{all\ x} g(x)P(X = x).$$

LOTUS says we can compute the expectation of $g(X)$ without knowing the PMF of $g(X)$.

Example 4.10. Compute $E(X)$ and $E(X^2)$ given the following distribution.

X	0	1	2
X^2	0	1	4
P	1/4	1/2	1/4

Solution: According to the distribution table, we compute the expectations as

$$E(X) = 0 \times 1/4 + 1 \times 1/2 + 2 \times 1/4 = 1;$$
$$E(X^2) = 0 \times 1/4 + 1 \times 1/2 + 4 \times 1/4 = 3/2.$$

Note that $E(X^2) \neq [E(X)]^2$.

Remark. In general, $E[g(X)] \neq g(E(X))$. Linearity implies that if g is a linear function of X, then $E[g(X)] = g(E(X))$. For a nonlinear function g, the relationship between $E[g(X)]$ and $g(E(X))$ is determined case by case. We will get back to this point when we learn Jensen's inequality.

Example 4.11 (St. Petersburg Paradox)**.** Flip a fair coin over and over again until the head lands the first time. You will win 2^k dollars if the head lands in the k-th trial (including the successful trial). What is the expected payoff of this game?

Solution: Let $X = 2^k$. We want to find $E(X)$. The probability of the first head showing up in the k-th trial is $\frac{1}{2^k}$. Therefore,

$$E(X) = \sum_{k=1}^{\infty} 2^k \cdot \frac{1}{2^k} = \sum_{k=1}^{\infty} 1 = \infty$$

The expected payoff is infinitely high! This is against most people's intuition. This is because we intuitively think that $E(X) = E(2^k) = 2^{E(k)}$, which is a finite number.

4.4 Median and mode

The mean is called a measure of *central tendency* because it tells us something about the center of a distribution, specifically its center of mass. Other measures of central tendency that are commonly used in statistics are the median and the mode, which we now define.

Definition 4.2. We say that c is a **median** of a random variable X if $P(X \leq c) \geq 1/2$ and $P(X \geq c) \geq 1/2$.

Definition 4.3. For a discrete random variable X, we say that c is a **mode** of X if it maximizes the PMF: $P(X = c) \geq P(X = x)$ for all x. For a continuous random variable X with PDF f, we say that c is a **mode** if it maximizes the PDF: $f(c) \geq f(x)$ for all x.

Intuitively, the median is a value c such that half the mass of the distribution falls on either side of c (or as close to half as possible, for discrete random variables), and the mode is a value that has the greatest mass or density out of all values in the support of X. If the CDF F is a continuous, strictly increasing function, then $F^{-1}(1/2)$ is the median (and is unique).

Remark. A distribution can have multiple medians and multiple modes. Medians have to occur side by side; modes can occur all over the distribution.

Example 4.12. The main reason why the median is sometimes preferred over the mean is that the median is more robust to extreme values. Think about an income distribution. Higher incomes are rare, but their absolute values are high. Thus, the mean income tends be higher than what the mass of the population would earn. But the median is more robust to extreme values and is closer to the earnings of an "average" person. For example, the mean of China's income is ¥$2,561$ monthly in 2019; the median is only ¥$2,210$.

Income (monthly, yuan)	<1k	1-2k	2-5k	5-10k	10-20k	>20k
Population (million)	550	420	360	63	7.8	0.7

Table 4.2: China monthly income per capita. Source: NBS 2019.

Theorem 4.4. *Let X be an random variable with mean μ , and let m be a median of X.*

- *The value of c that minimizes the mean squared error $E\left(X-c\right)^2$ is $c=\mu$.*
- *A value of c that minimizes the mean absolute error $E\left|X-c\right|$ is $c=m$.*

4.5 Variance and standard deviation

Expectation is the most commonly used summary of a distribution, as it indicates where values are likely centered. However, it provides limited insight into the distribution's overall shape. For example, two random variables might have the same mean, yet one could have values spread far from the mean while the other has values tightly clustered around it. Variance, on the other hand, describes how far values in a distribution typically deviate from the mean, offering a measure of the distribution's dispersion.

Definition 4.4. The **variance** of a random variable X is defined as

$$Var(X) = E\left(X - EX\right)^2.$$

The **standard deviation** of X is defined as

$$SD(X) = \sqrt{Var(X)}.$$

We often denote standard deviation by the greek letter $\sigma \equiv SD(X)$, and variance by σ^2.

Variance measures how far X typically deviates from its mean, but instead of averaging the differences, we average the squared differences to ensure both positive and negative deviations contribute. The expected deviation, $E(X - E(X))$, is always zero, so squaring avoids this cancellation. Since variance is in squared units, we take the square root to get the standard deviation, restoring the original units.

Why take squares? Sometimes we also use $E(|X - E(X)|)$ instead. But it is less common because the absolute value function isn't differentiable. Besides, squaring connects to geometric concepts like the distance formula and Pythagorean theorem, which have useful statistical meanings.

Theorem 4.5. *For any random variable X,*

$$Var(X) = E(X^2) - (EX)^2.$$

Proof. Let $\mu = E(X)$. By definition,

$$\begin{aligned} Var(X) &= E(X-\mu)^2 = E(X^2 - 2\mu X + \mu^2) \\ &= E(X^2) - 2\mu E(X) + \mu^2 = E(X^2) - \mu^2. \end{aligned}$$

□

Example 4.13. Find the variance for $X \sim \text{Bern}(p)$.

$$Var(X) = E(X^2) - E^2(X) = p - p^2 = p(1-p).$$

Theorem 4.6. *Variance has the following properties:*

- $Var(X) \geq 0$
- $Var(X + c) = Var(X)$
- $Var(cX) = c^2 Var(X)$
- *If* X, Y *are independent,* $Var(X + Y) = Var(X) + Var(Y)$.
- *If* $X_1, X_2, \ldots, X_n$ *are independent,* $Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i)$.

Example 4.14. Find the variance for $X \sim \text{Bin}(n, p)$. $X = X_1 + \cdots + X_n$ where X_i are *i.i.d* Bernoulli distributions

$$Var(X) \overset{iid}{=} \sum_{i=1}^{n} Var(X_i) = np(1 - p).$$

4.6 Covariance and correlation

For more than one random variable, it is also of interest to know the relationship between them. Are they dependent? How strong is the dependence? Covariance and correlation are intended to measure that dependence. But they only capture a particular type of dependence, namely linear dependence.

Definition 4.5. The **covariance** between random variables X and Y is defined as

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

The covariance between X and Y reflects how much X and Y simultaneously deviate from their respective means. If $X > EX, Y > EY$ or $X < EX, Y < EY$ simultaneously, then $Cov(X, Y)$ tends be positive. Conversely, if $X > EX$ is pair with $Y < EY$ (or $X < EX$ paired with $Y > EY$), then $Cov(X, Y)$ tends to be negative.

Theorem 4.7. *For any random variables* X *and* Y,

$$Cov(X, Y) = E(XY) - E(X)E(Y).$$

Proof. Let $\mu_X = E(X)$ and $\mu_Y = E(Y)$. By definition,

$$\begin{aligned} Cov(X,Y) &= E(XY - \mu_X Y - \mu_Y X + \mu_X \mu_Y) \\ &= E(XY) - \mu_X E(Y) - \mu_Y E(X) + \mu_X \mu_Y \\ &= E(XY) - E(X)E(Y). \end{aligned}$$

□

Theorem 4.8. *If X, Y are independent, they are uncorrelated. But the converse is false.*

Proof. $Cov(X,Y) = E(XY) - E(X)E(Y)$. Independence implies $E(XY) = E(X)E(Y)$. Thus, $Cov(X,Y) = 0$. But $Cov(X,Y) = 0$ does not necessarily imply independence. Consider the following counter example. Let X be a random variable that takes three values -1, 0, 1 with equal probability. And $Y = X^2$. X and Y are clearly dependent. But they their correlation is 0. Since $E(X) = 0$, $E(Y) = 2/3$, $E(XY) = E(X^3) = 0$, $Cov(X,Y) = 0$. □

Remark. Covariances and correlations provide measures of the extend to which two random variables are linearly related. If we plot the values of X and Y in the xy-plane, if the points form a straight line, that would signal a strong positive (if positive slope) or negative (if negative slope) correlation. It is possible that the correlation is 0 if X and Y are dependent but the relationship is nonlinear.

Theorem 4.9. *Covariance has the following properties:*

- $Cov(X,X) = Var(X)$
- $Cov(X,Y) = Cov(Y,X)$
- $Cov(cX,Y) = Cov(X,cY) = c\,[Cov(X,Y)]$
- $Cov(X+Y,Z) = Cov(X,Z) + Cov(Y,Z)$
- $Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$
- $Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i) + 2\sum_{i<j} Cov(X_i, X_j)$

Proof. We only prove the variance-covariance property:

$$\begin{aligned} Var(X+Y) &= E[(X+Y-\mu_X-\mu_Y)^2] \\ &= E[(X-\mu_X)^2+(Y-\mu_Y)^2+2(X-\mu_X)(Y-\mu_Y)] \\ &= Var(X)+Var(Y)+2Cov(X,Y). \end{aligned}$$

□

Exercise 4.1. Find $Cov(X+Y,Z+W)$ and $Var(X-Y)$.

While $\mathrm{Cov}(X,Y)$ quantifies how X and Y vary together, its magnitude also depends on the absolute scales of X and Y (multiply X by a constant c, the covariance will be different). To establish a measure of association between X and Y that is unaffected by arbitrary changes in the scales of either variable, we introduce a “standardized covariance” called correlation.

Definition 4.6. The **correlation** between random variables X and Y is defined as

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}.$$

We also denote correlation by $\rho \equiv Corr(X,Y)$.

Unlike covariance, scaling X or Y has no effect on the correlation. We can verify this:

$$Corr(cX,Y) = \frac{Cov(cX,Y)}{\sqrt{Var(cX)Var(Y)}} = \frac{cCov(X,Y)}{c\sqrt{Var(X)Var(Y)}} = Corr(X,Y).$$

Theorem 4.10. *For any random variable X and Y,*

$$-1 \le Corr(X,Y) \le 1.$$

Proof. Without loss of generality, assume X,Y both have variance 1, since scaling does not change the correlation. Let $\rho = Corr(X,Y) = Cov(X,Y)$. Then

$$\begin{aligned} Var(X+Y) = Var(X)+Var(Y)+2Cov(X,Y) = 2+2\rho \ge 0, \\ Var(X-Y) = Var(X)+Var(Y)-2Cov(X,Y) = 2-2\rho \ge 0. \end{aligned}$$

Thus $-1 \le \rho \le 1$. □

It is said that X and Y are **positively correlated** if $Corr(X, Y) > 0$, that X and Y are **negatively correlated** if $Corr(X, Y) < 0$, and that X and Y are **uncorrelated** if $Corr(X, Y) = 0$.

Theorem 4.11. *Suppose that X is a random variable and $Y = aX + b$ for some constants a, b, where $a \neq 0$. If $a > 0$, then $\rho_{XY} = 1$. If $a < 0$, then $\rho_{XY} = -1$.*

Proof. If $Y = aX+b$, then $E(Y) = aE(X)+b$. Thus, $Y-E(Y) = a(X-E(X))$. Therefore,

$$Cov(X, Y) = aE[(X - EX)^2] = aVar(X).$$

Since $Var(Y) = a^2Var(X)$, $\rho_{XY} = \frac{a}{|a|}$. The theorem thus follows. □

Example 4.15. Toss two coins. Let X be the number of Heads, and Y be the number of Tails. Find the covariance and correlation between X and Y.

Solution: Note that X and Y are counterparts to each other, $Y = 2 - X$. So we expect the correlation be negative. The expectation of X and Y are the same: $EX = EY = 1$. So we have $X - EX = -1, 0, 1$ and $Y - EY = 1, 0, -1$. The corresponding probabilities are 1/4,1/2,1/4 respectively. Therefore,

$$Cov(X, Y) = (-1) \times 1 \times 1/4 + 1 \times (-1) \times 1/4 = -1/2.$$

Since $Var(X) = Var(Y) = 1/2$, the correlation is

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}} = \frac{-1/2}{\sqrt{1/2 \times 1/2}} = -1.$$

Example 4.16. Let $X \sim HGeom(w, b, n)$. Find $Var(X)$.

Solution: Interpret X as the number of white balls in a sample of size n from an box with w white and b black balls. We can represent X as the sum of indicator variables, $X = I_1 + \cdots + I_n$, where I_j is the indicator of the j-th ball in the sample being white. Each I_j has mean $p = w/(w+b)$ and variance $p(1-p)$, but because the I_j are dependent, we cannot simply add their variances. Instead,

$$\begin{aligned} Var(X) &= Var\left(\sum_{j=1}^{n} I_j\right) \\ &= Var(I_1) + \cdots + Var(I_n) + 2\sum_{i<j} Cov(I_i, I_j) \\ &= np(1-p) + 2\binom{n}{2} Cov(I_i, I_j) \end{aligned}$$

In the last step, because of symmetry, for every pair i and j, $Cov(I_i, I_j)$ are the same.

$$\begin{aligned} Cov(I_i, I_j) &= E(I_i I_j) - E(I_i)E(I_j) \\ &= P(i \text{ and } j \text{ both white}) - P(i \text{ is white})P(j \text{ is white}) \\ &= \frac{w}{w+b} \cdot \frac{w-1}{w+b-1} - p^2 \\ &= p\frac{Np-1}{N-1} - p^2 \\ &= \frac{p(p-1)}{N-1} \end{aligned}$$

where $N = w + b$. Plugging this into the above formula and simplifying, we eventually obtain

$$Var(X) = np(1-p) + n(n-1)\frac{p(p-1)}{N-1} = \frac{N-n}{N-1}np(1-p).$$

This differs from the Binomial variance of $np(1-p)$ by a factor of $\frac{N-n}{N-1}$. This discrepancy arises because the Hypergeometric story involves sampling without replacement. As $N \to \infty$, it becomes extremely unlikely that we would draw the same ball more than once, so sampling with or without replacement essentially become the same.

Example 4.17 (PG exam)**.** Put k balls into n boxes. Let X be the number of empty boxes. Find $E(X)$ and $Var(X)$.

Solution: Define an indicator variable

$$I_j = \begin{cases} 1 & j\text{-th box is empty} \\ 0 & \text{otherwise} \end{cases}$$

Then $X = \sum_{j=1}^{n} I_j$. Unconditionally, the probability of one box being empty is $\left(\frac{n-1}{n}\right)^k$. Therefore,

$$E(I_j) = P(j\text{-th box is empty}) = \left(\frac{n-1}{n}\right)^k$$

for $j = 1, 2, \ldots, n$. It follows that

$$E(X) = \sum_{j=1}^{n} I_j = nE(I_j) = n\left(\frac{n-1}{n}\right)^k.$$

To compute the variance,

$$\begin{aligned} Var(X) = Var(I_1 + \cdots + I_n) &= \sum_{j=1}^{n} Var(I_j) + 2\sum_{i<j} Cov(I_i, I_j) \\ &= nVar(I_j) + 2\binom{n}{2} Cov(I_i, I_j), \end{aligned}$$

since by symmetry, $Var(I_j)$ is the same for all j and $Cov(I_i, I_j)$ is the same for all $i \neq j$. It suffices to compute $Var(I_j)$ and $Cov(I_i, I_j)$ for any j and $i \neq j$. Since I_j only takes number 0 and 1,

$$E(I_j^2) = \left(\frac{n-1}{n}\right)^k,$$

$$Var(I_j) = E(I_j^2) - (E(I_j))^2 = \left(\frac{n-1}{n}\right)^k - \left(\frac{n-1}{n}\right)^{2k}.$$

For the covariance term,

$$E(I_i I_j) = P(i, j \text{ are both empty}) = \left(\frac{n-2}{n}\right)^k,$$

$$Cov(I_i, I_j) = E(I_i I_j) - E(I_i)E(I_j) = \left(\frac{n-2}{n}\right)^k - \left(\frac{n-1}{n}\right)^{2k}.$$

Therefore,

$$Var(X) = n\left[\left(\frac{n-1}{n}\right)^k - \left(\frac{n-1}{n}\right)^{2k}\right] + 2\binom{n}{2}\left[\left(\frac{n-2}{n}\right)^k - \left(\frac{n-1}{n}\right)^{2k}\right].$$

4.7 Moments and MGF

Definition 4.7. Let X be a random variable with mean μ and variance σ^2 . For any positive integer n, the n-th **moment** of X is $E(X^n)$, the n-th **central moment** is $E(X-\mu)^n$, and the n-th **standardized moment** is $E\left(\frac{X-\mu}{\sigma}\right)^n$.

In accordance with this terminology, $E(X)$ is the first moment of X, $Var(X)$ is the second central moment of X. It is natural to ask if there are higher order moments. The answer is yes.

Definition 4.8. Let X be a random variable with mean μ, standard deviation σ, and finite third moment. The **skewness** of X is defined as

$$\text{Skew}(X) = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right].$$

Definition 4.9. The **Kurtosis** of X is defined as

$$\text{Kurt}(X) = \left[\left(\frac{X-\mu}{\sigma}\right)^4\right].$$

Skewness is the measure of the lopsidedness of the distribution; any symmetric distribution will have a third central moment, if defined, of zero. A distribution that is skewed to the left (the tail of the distribution is longer on the left) will have a negative skewness. A distribution that is skewed to the right (the tail of the distribution is longer on the right), will have a positive skewness.

Kurtosis is a measure of the heaviness of the tail of the distribution. If a distribution has heavy tails, the kurtosis will be high; conversely, light-tailed distributions have low kurtosis.

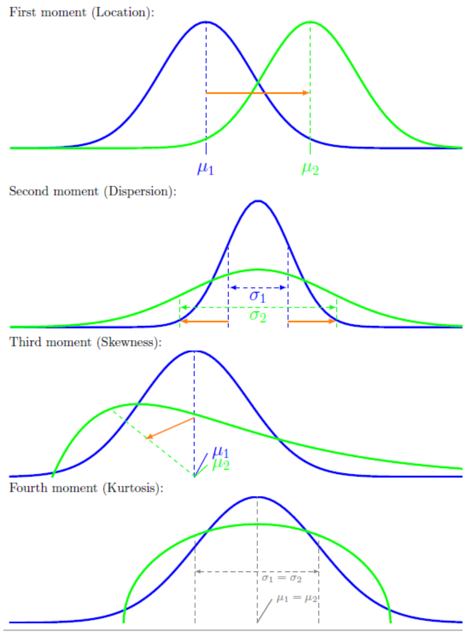


Figure 4.1: Moments and the shape of a distribution

We see that moments give information about the shape of a distribution. Different orders of moments captures different aspects of the distribution. In fact, if we know all the moments (moments of infinitely high order), we can exactly pin down the distribution.

Theorem 4.12. *For a distribution of mass or probability on a bounded interval, the collection of all the moments (of all orders, from* 0 *to* ∞*) uniquely determines the distribution.*

So there are two ways of fully characterize a distribution:

1. Listing all the possible values along with their associated probabilities;
2. Giving all the moments of the distribution.

It is somewhat like the analogous Taylor theorem in the probability theory. We can represent any distribution by a sequence of higher order "polynomials": $E(X), E(X^2), E(X^3), \dots$

Definition 4.10. Let X be a random variable. For each real number t, define the **moment generating function** (MGF) as

$$M_X(t) = E\left(e^{tX}\right).$$

To see why it is "generating" moments, take the Taylor expansion of the exponential function:

$$e^{tX} = 1 + tX + \frac{t^2X^2}{2!} + \frac{t^3X^3}{3!} + \cdots$$

Hence,

$$M_X(t) = E\left(e^{tX}\right) = 1 + E(X)t + E(X^2)\frac{t^2}{2!} + \cdots$$

A natural question at this point is: What is the interpretation of t? The answer is that t has no interpretation in particular; it's just a bookkeeping device that we introduce in order to *encode* the sequence of moments in a differentiable function.

Theorem 4.13. *Let $M_X(t)$ be the MGF of X. Then the n-th moment of X is given by $E(X^n) = M_X^{(n)}(0)$, where $M_X^{(n)}$ denotes the n-th derivative of the MGF.*

Theorem 4.14. *The MGF of a random variable determines its distribution: if two random variables have the same MGF, they must have the same distribution.*

Theorem 4.15. *If X and Y are independent, then the MGF of $X+Y$ is the product of the individual MGFs:*

$$M_{X+Y}(t) = M_X(t)M_Y(t).$$

Example 4.18. For $X \sim Bern(p)$, e^{tX} takes on the value e^t with probability p and the value 1 with probability q, so $M(t) = E\left(e^{tX}\right) = pe^t + q$. Since this is finite for all values of t, the MGF is defined on the entire real line.

Example 4.19. The MGF of a $Bin(n, p)$ random variable is $M(t) = (pe^t + q)^n$, since it is the product of n independent Bernoulli MGFs.

4.8 Poisson distribution

Now we introduce arguably the most popular discrete distribution—Poisson distribution. Poisson distribution is used to model independent events occurring at a constant mean rate. It is like the Binomial distribution in the sense that they both model the number of occurrence of events, but it is parametrized on the "rate" of the event (how many times an event occurs in a unit of time on average) rather than the total number of events and the probability of each event. It is therefore more practical in real-world modeling since we mostly observe the rate rather than the totality. We introduce the Poisson distribution by showing that it is a limiting case of the Binomial distribution.

Problem 4.1. Suppose we are studying the distribution of the number of visitors to a certain website. Every day, a million people independently decide whether to visit the site, with probability $p = 2 \times 10^{-6}$ of visiting. What is the probability of getting k visitors on a particular day?

We can model the problem with a Binomial distribution. Let $X \sim Bin(n, p)$ be the number of visitors, where $n = 10^6$ and $p = 2 \times 10^{-6}$. But it is easy to run into computational difficulties with such a large n and small p. This is not uncommon, if we want to model the number of emails one receives per day, or the number of phone calls in a service center. In such cases, we could reasonably assume $n \to \infty$ and $p \to 0$ while $np = \lambda$ is a constant. We may call λ — the "rate", as it can be interpreted as the average visitors per day.

Take limit of the Binomial distribution:

$$\begin{aligned}
P(X = k) &= \lim_{n\to\infty} \binom{n}{k} p^k (1-p)^{n-k} \\
&= \lim_{n\to\infty} \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\
&= \lim_{n\to\infty} \frac{n!}{(n-k)!k!} \frac{\lambda^k}{n^k} \underbrace{\left(1 - \frac{\lambda}{n}\right)^n}_{\to e^{-\lambda}} \underbrace{\left(1 - \frac{\lambda}{n}\right)^{-k}}_{\to 1} \\
&= \lim_{n\to\infty} \underbrace{\frac{n!}{n^k (n-k)!}}_{\to 1} \frac{\lambda^k}{k!} e^{-\lambda} \\
&= \frac{\lambda^k}{k!} e^{-\lambda}.
\end{aligned}$$

This is the PMF of the Poisson distribution.

Definition 4.11. A random variable X has the **Poisson distribution** with parameter λ if the PMF of X is

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots$$

We denote this as $X \sim \text{Pois}(\lambda)$. We can easily verify this is a valid PMF because $\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{\lambda}$.

Theorem 4.16. *If $X \sim Bin(n, p)$ and we let $n \to \infty$ and $p \to 0$ such that $\lambda = np$ remains fixed, then the PMF of X converges to the PMF of $Pois(\lambda)$.*

The expectation of the Poisson distribution is

$$\begin{aligned} E(X) =& \sum_{k=0}^{\infty} k \cdot \frac{e^{-\lambda}\lambda^k}{k!} \\ =& e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} \\ =& \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \\ =& \lambda e^{-\lambda} e^{\lambda} = \lambda. \end{aligned}$$

To get the variance, we first compute $E(X^2)$. By LOTUS,

$$E(X^2) = \sum_{k=0}^{\infty} k^2 \cdot \frac{e^{-\lambda}\lambda^k}{k!} = e^{-\lambda} \sum_{k=1}^{\infty} k^2 \frac{\lambda^k}{k!}$$

Differentiate $\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{\lambda}$ on both sides with respect to λ and multiply (replenish) again by λ:

$$\sum_{k-1}^{\infty} k \frac{\lambda^k}{k!} = \lambda e^{\lambda}$$

Repeat:

$$\sum_{k-1}^{\infty} k^2 \frac{\lambda^k}{k!} = \lambda(e^{\lambda} + \lambda e^{\lambda})$$

Therefore, we have

$$E(X^2) = e^{-\lambda}(\lambda + \lambda^2)e^{\lambda} = \lambda + \lambda^2$$

Finally,

$$Var(X) = E(X^2) - (E(X))^2 = \lambda + \lambda^2 - \lambda^2 = \lambda.$$

Example 4.20. Continued with the website visiting example, there are one million people visiting the site every day, each with probability $p = 2 \times 10^{-6}$. Give an approximation for the probability of getting at least three visitors on a particular day.

Let X be the number of visitors. Since n is large, p is small, $np = 2$ is fixed, X is well approximated by $Pois(2)$. Therefore,

$$\begin{aligned} P(X \geq 3) = 1 - P(X < 3) &= 1 - P(X = 0) - P(X = 1) - P(X = 2) \\ &= 1 - e^{-2} - 2e^{-2} - \frac{2^2}{2!}e^{-2} \\ &= 1 - 5e^{-2} \approx 0.32. \end{aligned}$$

Example 4.21. What is the probability of an earthquakes in a year in Sichuan?

Historical records[2] show that, from 26 BCE to 2021 CE, there were 309 earthquakes with magnitude of 5.0 or greater. Let X be the number of earthquakes with magnitude 5.0 or greater. The annual rate λ of earthquakes is therefore $\frac{309}{2048} = 0.15$. Assume earthquakes are independent events (not always the case). Then $X \sim \text{Pois}(0.15)$. By the distribution of the Poisson distribution,

$$P(X = k) = \begin{cases} 0.86 & k = 0 \\ 0.13 & k = 1 \\ 0.01 & k = 2 \end{cases}.$$

The Poisson distribution is often used in situations where we are counting the number of successes in a particular region or interval of time, where there are a large number of trials, each with a small probability of success. The Poisson paradigm says in situations like this, we can approximate the number of successes by a Poisson distribution. It is more general than Theorem 4.16, as we relax the assumption of independence and identical events.

[2] See this article from the Sichuan Earthquake Administration.

Proposition 4.1 (Poisson paradigm)**.** *Let $A_1, \dots, A_n$ be events with $p_j = P(A_j)$, where n is large, the p_j are small, and the A_j are independent or weakly dependent. Then $X = \sum_{j=1}^{n} I(A_j)$, that is how many of the A_j occur, is approximately distributed as $Pois(\lambda)$ with $\lambda = \sum_{j=1}^{n} p_j$.*

The Poisson paradigm is also called the *law of rare events*. The interpretation of "rare" is that the p_j are small, but λ is relatively stable. The number of events that occur may not be exactly Poisson, but the Poisson distribution often gives good approximations. Note that the conditions for the Poisson paradigm to hold are fairly flexible: the n trials can have different success probabilities, and the trials don't have to be independent, though they should not be very dependent. So there are a wide variety of situations that can be cast in terms of the Poisson paradigm. This makes the Poisson a very popular model.

Example 4.22. If we have m people and $\binom{m}{2}$ pairs. Each pair of people has probability $p = 1/365$ of having the same birthday. Find the probability of at least one match.

Solution: The probability of match is small, and the number of pairs is large. We consider using the Poisson paradigm to approximate the number X of birthday matches. $X \approx Pois(\lambda)$ where $\lambda = \binom{m}{2}\frac{1}{365}$. Then the probability of at least one match is

$$P(X \geq 1) = 1 - P(X = 0) \approx 1 - e^{-\lambda}.$$

For $m = 23$, $\lambda = 253/365$ and $1 - e^{-\lambda} \approx 0.5$, which agrees with our previous finding that we need 23 people to have 50% chance of a birthday match.

Example 4.23. Continued with the assumption above. What's the probability of two people who were born not only on the same day, but also at the same hour and the same minute?

Solution: This is the birthday problem with $c = 365 \cdot 24 \cdot 60 = 525600$ categories rather than 365 categories. By Poisson approximation, the probability of at least one match is approximately $1 - e^{-\lambda_1}$ where $\lambda_1 = \binom{m}{2}\frac{1}{525600}$. This would require $m = 854$ to reach the break even point, 50% chance of getting a match.

Theorem 4.17. *If $X \sim Pois(\lambda_1)$, $Y \sim Pois(\lambda_2)$, and X, Y are independent, then $X + Y \sim Pois(\lambda_1 + \lambda_2)$.*

Proof. To get the PMF of $X + Y$, condition on X and use the law of total probability:

$$\begin{aligned}
P(X+Y=k) &= \sum_{j=0}^{k} P(X+Y=k|X=j)P(X=j) \\
&= \sum_{j=0}^{k} P(Y=k-j)P(X=j) \\
&= \sum_{j=0}^{k} \frac{e^{-\lambda_2}\lambda_2^{k-j}}{(k-j)!} \cdot \frac{e^{-\lambda_1}\lambda_1^{j}}{j!} \\
&= \frac{e^{-(\lambda_1+\lambda_2)}}{k!} \sum_{j=0}^{k} \binom{k}{j} \lambda_1^j \lambda_2^{k-j} \\
&= \frac{e^{-(\lambda_1+\lambda_2)}}{k!} (\lambda_1+\lambda_2)^k.
\end{aligned}$$

We thus arrive at the PMF for $Pois(\lambda_1+\lambda_2)$. The intuition is, if there are two different types of events occurring at rates λ_1 and λ_2, independently, then the overall event rate is $\lambda_1+\lambda_2$. □

4.9 Inequalities*

This section introduces some of the most popular inequality in statistics and general mathematics. Interestingly, our probability theories can shed light on these inequalities that are otherwise hard to explain. We don't show formal proofs here, but just point out how these inequalities can be useful in statistics.

Theorem 4.18 (Cauchy-Schwarz inequality)**.**

$$\left|\sum x_i y_i\right| \le \sqrt{\sum x_i^2}\sqrt{\sum y_i^2}$$

Proof. If X, Y have zero means, their correlation can be written as

$$\rho_{XY} = \frac{E(XY)}{\sqrt{E(X^2)E(Y^2)}}$$

Since $|\rho_{XY}| \le 1$, we always have

$$|E(XY)| \le \sqrt{E(X^2)E(Y^2)}.$$

Consider $\{x_i\}$ and $\{y_i\}$ as realizations of X and Y with equal probabilities, such that $E(X) = \frac{1}{n}\sum x_i$. The original inequality is thus proved. □

Theorem 4.19 (Jensen's inequality)**.** *For a convex function f, we have*

$$\frac{1}{n}\sum f(x_i) \geq f\left(\frac{1}{n}\sum x_i\right);$$

If f is concave, then

$$\frac{1}{n}\sum f(x_i) \leq f\left(\frac{1}{n}\sum x_i\right).$$

Proof. This is not a proof, but a special case that helps to understand Jensen's inequality. Since

$$Var(X) = E(X^2) - (E(X))^2 \geq 0$$

We have

$$E(X^2) \geq (E(X))^2.$$

Note that $f(X) = X^2$ is a convex function, and $E(*) = \frac{1}{n}\sum *$, we have shown the first inequality. The concave case is the opposite.

In general, if g is a convex function, then $E(g(X)) \geq g(E(X))$. If g is a concave function, then $E(g(X)) \leq g(E(X))$. In both cases, the only way that equality can hold is if there are constants a and b such that $g(X) = a + bX$ with probability 1. □

Theorem 4.20 (Markov inequality)**.** *Let X be a random variable, then*

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

That is, the probability of $|X|$ deviating from its mean by a multiple of a must be less than $1/a$.

Proof. Define a random variable

$$I_{|X|\geq a} = \begin{cases} 1 & \text{if } |X| \geq a \\ 0 & \text{if } |X| < a \end{cases}$$

Note that $P(|X| \geq a) = E(I_{|X|\geq a})$. It always holds that

$$a \cdot I_{|X|\geq a} \leq |X|$$

Therefore,

$$E\left[a \cdot I_{|X|\geq a}\right] \leq E|X|$$

Hence,

$$P(|X| \geq a) \leq \frac{E|X|}{a}.$$

□

For an intuitive interpretation, let X be the income of a randomly selected individual from a population. Taking $a = 2E(X)$, Markov's inequality says that $P(X \geq 2E(X)) \leq 1/2$, i.e., it is impossible for more than half the population to make at least twice the average income. This is clearly true, since if over half the population were earning at least twice the average income, the average income would be higher. Similarly, $P(X \geq 3E(X)) \leq 1/3$: you can't have more than 1/3 of the population making at least three times the average income, since those people would already drive the average above what it is.

Theorem 4.21 (Chebyshev inequality)**.** *Let X be a random variable with mean μ and standard deviation σ, then*

$$P\left(|X - \mu| > c\sigma\right) \leq \frac{1}{c^2}$$

That is, the probability of X deviating from its mean by a times the standard deviation must be less than $1/a^2$.

Proof. We first show

$$P(|X - \mu| > a) \leq \frac{\sigma^2}{a^2}$$

This is true by taking squares and applying the Markov inequality,

$$P(|X-\mu|>a)=P((X-\mu)^2>a^2)\le\frac{E(X-\mu)^2}{a^2}=\frac{\sigma^2}{a^2}.$$

Substitute $c\sigma$ for a, we have the original inequality. □

This gives us an upper bound on the probability of a random variable being more than c standard deviations away from its mean, e.g., there can't be more than a 25% chance of being 2 or more standard deviations from the mean. Given the mean and standard deviation of a random variable X, we know that $\mu\pm2\sigma$ captures 75% of its possible values; $\mu\pm3\sigma$ captures 90% of the possible values.

Chapter 5

Continuous Distributions

5.1 Continuous random variables

Continuous random variables, in many ways, are more versatile and useful than discrete distributions. One key reason is that many quantities in the physical world, such as temperature, height, weight, and time, are inherently continuous in nature. These variables can take on any value within a range, providing a more accurate representation of real-world phenomena compared to discrete variables, which are limited to distinct values. Additionally, the probability density functions (PDFs) of continuous distributions are often defined by smooth, differentiable functions. This mathematical structure allows us to apply calculus for analysis, enabling precise calculations of probabilities, expected values, and other statistical measures. The ability to integrate and differentiate these functions not only simplifies manipulation but also makes continuous distributions a powerful tool for solving complex problems in physics, engineering, and data analysis.

Definition 5.1. A random variable has a continuous distribution if its CDF is *differentiable*. A continuous random variable is a random variable with a continuous distribution.

Definition 5.2. For a continuous random variable X with CDF F, the **probability density function** (PDF) of X is the derivative of the CDF, given by $f(x) = F'(x)$. The support of X is the set of all x where $f(x) > 0$.

Remark. By the fundamental theorem of calculus, we integrate a PDF to get the CDF:

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

PDF differs from the discrete PMF in important ways:

- For a continuous random variable, $P(X = x) = 0$ for all x;
- The quantity $f(x)$ is not a probability. To get the probability, we integrate the PDF (probability is the area under the PDF):

$$P(a < X \leq b) = F(b) - F(a) = \int_{a}^{b} f(x)dx.$$

- Since any single value has probability 0, including or excluding endpoints does not matter.

$$P(a < X < b) = P(a < X \leq b) = P(a \leq X < b) = P(a \leq X \leq b).$$

Theorem 5.1. *The PDF f of a continuous random variable must satisfy the following criteria:*

- *Nonnegative: $f(x) \geq 0$;*
- *Integrates to 1: $\int_{-\infty}^{\infty} f(x)dx = 1$.*

Definition 5.3. The **expectation** of a continuous random variable X with PDF f is

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

Theorem 5.2. *If X is a continuous random variable with PDF f and $g : \mathbb{R} \to \mathbb{R}$. The LOTUS applies*

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

	Discrete	Continuous
PMF/PDF	$P(X=x)=p(x)$	$P(a\leq X\leq b)=\int_a^b f(x)dx$
CDF	$F(x)=P(X\leq x)=\sum_{k\leq x} p(k)$	$F(x)=P(X\leq x)=\int_{-\infty}^x f(t)dt$
Expectation	$E(x)=\sum_x xP(X=x)$	$E(X)=\int_{-\infty}^{+\infty} xf(x)dx$
LOTUS	$E[g(x)]=\sum_x g(x)P(X=x)$	$E[g(x)]=\int_{-\infty}^{+\infty} g(x)f(x)dx$

5.2 Special integrals

There are many reasons to learn integrals. But the most compelling reason is that math is no longer the same with integrals. We can have many amazing results with integrals that were otherwise not imaginable. This section is not directly related to our main theme. But let's take a detour just to appreciate the beauty of integrals.

Example 5.1. Show that $\int_{-\infty}^{+\infty} e^{-x^2}dx=\sqrt{\pi}$.

Proof. This is known as Gaussian integral, which is the kernel of the PDF of the normal distribution. It also amazingly relates two of the most famous constants in mathematics. It is not integrable by normal integration techniques. But it can be solved by switching to the polar coordinate.

$$\begin{aligned}
\left(\int_{-\infty}^{+\infty} e^{-x^2}dx\right)^2 &= \int_{-\infty}^{+\infty} e^{-x^2}dx\int_{-\infty}^{+\infty} e^{-y^2}dy \\
&= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} e^{-(x^2+y^2)}dxdy \\
&= \int_0^{2\pi}\int_0^{\infty} e^{-r^2}rdrd\theta \qquad\qquad dA=dxdy=rdrd\theta \\
&= \int_0^{2\pi}\int_0^{\infty}\frac{1}{2}e^{-u}dud\theta \qquad\qquad \text{let } u=r^2 \\
&= \frac{1}{2}\int_0^{2\pi} d\theta=\pi.
\end{aligned}$$

□

Example 5.2. Show that $\int_0^\infty t^n e^{-t} dt = n!$

Proof. $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$ is known as the Gamma function, which is definitely one of the most interesting functions in mathematics. It is the extension of factorials to real numbers or even complex numbers. It also has many interesting properties, such as $\Gamma(n) = (n-1)!$, $\Gamma(1/2) = \sqrt{\pi}$, $\Gamma(3/2) = \sqrt{\pi}/2$, $\Gamma'(1) = -\gamma$ and so on. The $(n-1)$ in the Gamma function is due to historical reasons and does not matter in our case. We will prove the integral with n instead of $(n-1)$.

There are many ways to prove this. One is to discover the recursive relationship $\Gamma(n+1) = n\Gamma(n)$. But it does not give a clue why we need this integral to approximate the factorial. We start with an elementary integral

$$\int_0^\infty e^{at} dt = -\frac{1}{a}$$

where $a < 0$. Differentiate both sides n times with respect to a:

$$\begin{aligned}
\int_0^\infty e^{at} t dt = & \quad -(-1)a^{-2} \\
\int_0^\infty e^{at} t^2 dt = & \quad -(-1)(-2)a^{-3} \\
\int_0^\infty e^{at} t^3 dt = & \quad -(-1)(-2)(-3)a^{-4} \\
& \vdots \\
\int_0^\infty e^{at} t^n dt = & \quad (-1)^{n+1} n! a^{-(n+1)}
\end{aligned}$$

Let $a = -1$, we have

$$\int_0^\infty e^t t^n = n!$$

□

5.3 Uniform distribution

Definition 5.4. Let a and b be two given real numbers such that $a < b$. Let X be a random variable such that it is known that $a \leq X \leq b$ and, for every subinterval of $[a, b]$, the probability that X will belong to that subinterval is proportional to the length of that subinterval. We then say that the random variable X has the **Uniform distribution** on the interval $[a, b]$. The PDF of X is

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

This is a valid PDF since

$$\int_{-\infty}^{+\infty} f(x)dx = \int_a^b \frac{1}{b-a}dx = \frac{1}{b-a}\int_a^b dx = 1.$$

The CDF of X is

$$F(x) = \int_{-\infty}^{x} f(t)dt = \int_a^x f(t)dt = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \leq x \leq b \\ 1 & x > b \end{cases}.$$

The expectation of X:

$$E(X) = \int_a^b x\frac{1}{b-a}dx = \frac{1}{b-a}\left[\frac{x^2}{2}\right]_a^b = \frac{a+b}{2}.$$

To figure out the variance, first compute

$$E(X^2) = \int_a^b x^2\frac{1}{b-a}dx = \frac{1}{b-a}\left[\frac{x^3}{3}\right]_a^b = \frac{a^2+ab+b^2}{3}$$

Thus,

$$Var(X) = E(X^2) - E^2(X) = \frac{a^2+ab+b^2}{3} - \frac{(a+b)^2}{4} = \frac{(b-a)^2}{12}.$$

Exercise 5.1. A stick of unit length is broken at a random point X. What is the expected length of the longer piece?

Solution: The lengths of the two pieces are X and $1-X$, with $X \sim Unif(0,1)$. The longer piece is $\max(X, 1-X)$. For $X < 0.5$, the longer piece is $1-X$, and for $X \geq 0.5$, it is X. The expected value is:

$$E[\max(X, 1-X)] = \int_0^{0.5} (1-X)\,dx + \int_{0.5}^1 X\,dx = \frac{3}{4}.$$

5.4 Normal distribution

The most widely used model for random variables with continuous distributions is the family of normal distributions. One reason is that many real world samples appears to be normally distributed (the mass centered around the mean). The other reason is because of the Central Limit Theorem (will be discussed in later chapters), which essentially says the sum (or mean) or any random samples are approximately normal.

Definition 5.5. A random variable Z has the **standard Normal distribution** with mean 0 and variance 1, denoted as $Z \sim N(0,1)$, if Z has a PDF that follows

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$

This is a valid PDF because $\int_{-\infty}^{\infty} f(z)dz = 1$, which directly follows from Example 5.1. We further verify its mean and variance:

$$E(Z) = \int_{-\infty}^{+\infty} z \cdot \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = 0 \quad \text{by symmetry.}$$

$$\begin{aligned} Var(Z) &= E(Z^2) - (EZ)^2 = E(Z^2) \\ &= \int_{-\infty}^{+\infty} z^2 \cdot \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \\ &= \frac{2}{\sqrt{2\pi}} \int_0^{\infty} \underbrace{z}_{u} \cdot \underbrace{ze^{-z^2/2}dz}_{dv} \\ &= \frac{2}{\sqrt{2\pi}} \left\{ \left[z(-e^{-z^2/2}) \right]_0^{\infty} + \underbrace{\int_0^{\infty} e^{-z^2/2} dz}_{\sqrt{2\pi}/2} \right\} \\ &= 1. \end{aligned}$$

Definition 5.6. The CDF of standard normal distribution is usually denoted by Φ. Therefore,

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2} dt.$$

By symmetry, we have $\Phi(-z) = 1 - \Phi(z)$.

Definition 5.7. Let $X = \mu + \sigma Z$ where $Z \sim N(0, 1)$. Then we say X has the **Normal distribution** with mean μ and variance σ^2, denoted as $X \sim N(\mu, \sigma^2)$. The PDF of X is given by

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

The mean and variance of X can be easily verified by the properties of expectation and variance.

$$\begin{aligned} E(X) &= E(\mu + \sigma Z) = \mu + \sigma E(Z) = \mu, \\ Var(X) &= Var(\mu + \sigma Z) = \sigma^2 Var(Z) = \sigma^2. \end{aligned}$$

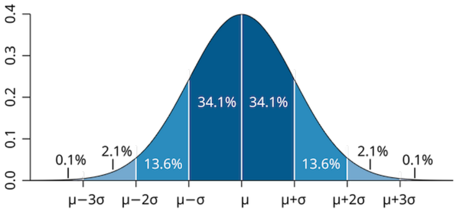
To verify the PDF, we utilize the standard normal CDF:

$$P(X \leq x) = P\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

The PDF is the derivative of the CDF,

$$f(x) = \frac{1}{\sigma}\Phi'\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right].$$

The shape of the normal distribution is the famous bell-shaped curve.



The normal distribution has the "three-sigma rule":

$$
\begin{aligned}
P(|X-\mu| \leq \sigma) &\approx 0.68 \\
P(|X-\mu| \leq 2\sigma) &\approx 0.95 \\
P(|X-\mu| \leq 3\sigma) &\approx 0.997
\end{aligned}
$$

Critical values: $\Phi(-1) \approx 0.16, \Phi(-2) \approx 0.025, \Phi(-3) \approx 0.0015$.

Theorem 5.3. *Let X have the Normal distribution with mean μ and variance σ^2. Let F be the CDF of X. Then the **standardization** of X*

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

has the standard normal distribution, and, for all x

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

To find the value of $\Phi(z)$, we need to use the normal probability table or statistical softwares.

Example 5.3. Suppose the test score of a class of 50 students is normally distributed with mean 80 and standard deviation 20 (the total mark is 100). A student has scored 90. What is his percentile in the class?

Solution: $X \sim N(80, 20)$. We want to find $P(X < 90)$. Standardize the distribution

$$P(X < 90) = P\left(\frac{X-80}{20} < \frac{90-80}{20}\right) = \Phi(0.5) \approx 0.69.$$

Theorem 5.4. *Suppose $X \sim N(\mu, \sigma^2)$. If $Y = aX + b$, then Y has the Normal distribution $Y \sim N(a\mu + b, a^2\sigma^2)$.*

Theorem 5.5. *If the random variables $X_1, \ldots, X_k$ are independent and $X_i \sim N(\mu_i, \sigma_i^2)$. Then*

$$X_1 + \cdots + X_k \sim N(\mu_1 + \cdots + \mu_k, \sigma_1^2 + \cdots + \sigma_k^2).$$

Example 5.4. Suppose the heights (in inches) of women and men independently follow the normal distribution, $X \sim N(165, 25)$, $Y \sim N(170, 25)$. Determine the probability that a randomly selected woman will be taller than a man.

Solution: Let $W = Y - X \sim N(170 - 165, 25 + 25)$. Then $W \sim N(5, 50)$. Therefore,

$$P(W < 0) = P\left(\frac{W-5}{\sqrt{50}} < \frac{-5}{\sqrt{50}}\right) = P\left(Z < -\frac{1}{\sqrt{2}}\right) = \Phi(-0.707) \approx 0.24.$$

Example 5.5 (Distribution of sample mean)**.** Let X_i be the height of a random individual. Assume $X_i \sim N(\mu, \sigma^2)$. Let $\{X_1, X_2, \ldots, X_n\}$ be a sample of n people. The sample mean is calculated as $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Determine the mean and variance of $\overline{X}_n$.

Solution: By the theorem above, the sum of a series of normal distributions is also normal:

$$\sum_{i=1}^{n} X_i = nX_i \sim N(n\mu, n\sigma^2)$$

since we assume all X_i follow the same distribution. Therefore, the distribution of the sample mean is

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \sigma^2/n).$$

That is, $\bar{X}_n$ has the normal distribution with mean μ and variance σ^2/n.

How do we understand the sample mean is also a random variable? A sample is a collection of random variables (each observation is a random variable in the sense that the outcome is uncertain). If you were to choose another sample, you would have a different sample mean. Therefore, the sample mean is also a random variable.

5.5 Chi-Square and Student-t*

We now introduce two distributions that are closely related to the Normal distribution.

Definition 5.8. Let $V = Z_1^2 + \cdots + Z_n^2$ where $Z_1, Z_2, \ldots, Z_n$ are i.i.d $N(0,1)$. Then V is said to have the **Chi-Square distribution** with n degrees of freedom, denoted as $V \sim \chi^2(n)$.

The χ^2 distribution is a special case of the Gamma distribution that will be introduced in the following sections. In fact, $\chi^2(1)$ is Gamma$(\frac{1}{2}, \frac{1}{2})$; $\chi^2(n)$ is Gamma$(\frac{n}{2}, \frac{1}{2})$.

Definition 5.9. Let

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

where $Z \sim N(0,1)$, $V \sim \chi^2(n)$, and Z is independent of V. Then T is said to have the **Student-*t* distribution** with n degrees of freedom, denoted as $T \sim t_n$.

Student-t distribution is symmetric and has the similar bell-shaped curve of the Normal distribution but with heavier tail. As $n \to \infty$, t_n distribution approaches the standard Normal distribution.

5.6 Exponential distribution

Imagine you are a shop owner that waits for your next customer. The customers arrive randomly, with no preference for any specific time interval. What interests us is the waiting time until the next customer arrives. Since the customers arrives randomly, the likelihood of it coming in the next moment is the same whether you've been waiting for one minute or ten minutes. In other words, the waiting time between events that occur randomly and independently over time. The exponential distribution is the mathematical model that best describes such scenarios.

To model the waiting time, let X represent the time until the next event. A crucial feature of this process is that the waiting time has no "memory." That is, no matter how long you've already waited, the probability of waiting an additional amount of time is the same. Mathematically, this memoryless property is expressed as:

$$P(X \geq s + t \mid X \geq s) = P(X \geq t), \quad \text{for all } s, t \geq 0.$$

The conditional probability can be rewritten using the definition of conditional probabilities:

$$P(X \geq s+t \mid X \geq s) = \frac{P(X \geq s+t)}{P(X \geq s)}.$$

Thus, the memoryless property implies:

$$\frac{P(X \geq s+t)}{P(X \geq s)} = P(X \geq t).$$

Let the survival function $S(x)$ represent $P(X \geq x)$. Substituting $S(x)$ into the equation gives:

$$\frac{S(s+t)}{S(s)} = S(t).$$

This reminds us of the exponential function. In fact, the only continuous and non-negative solution to this equation is:

$$S(x) = e^{-\lambda x}, \quad \lambda > 0,$$

where λ is a positive constant. This solution represents the probability that the waiting time exceeds x , and λ determines how quickly the probability decreases over time.

The CDF of X is exactly the opposite of $S(x)$:

$$F(x) = 1 - S(x) = 1 - e^{-\lambda x}.$$

Take derivative to get the PDF:

$$f(x) = F'(x) = \lambda e^{-\lambda x}.$$

Definition 5.10. A random variable X is said to have the **Exponential distribution** with parameter λ if its PDF is

$$f(x) = \lambda e^{-\lambda x}, \qquad x > 0.$$

We denote this as $X \sim \text{Expo}(\lambda).\lambda$ is interpreted as the "rate", i.e. number of events per unit of time.

To compute the expectation and variance, we first standardize the exponential distribution. Let $Y = \lambda X$, then $Y \sim$ Expo(1), because

$$P(Y \leq y) = P(X \leq y/\lambda) = 1 - e^{-y}.$$

It follows that,

$$\begin{aligned} E(Y) &= \int_0^\infty ye^{-y}dy = \left[-ye^{-y}\right]_0^\infty + \int_0^\infty e^{-y}dy = 1; \\ Var(Y) = E(Y^2) - (EY)^2 &= \int_0^\infty y^2e^{-y}dy - 1 = 1. \end{aligned}$$

For $X = Y/\lambda$, we have $E(X) = \frac{1}{\lambda}$, $Var(X) = \frac{1}{\lambda^2}$.

Theorem 5.6 (Memoryless property)**.** *If X has the exponential distribution with parameter λ, and let $t > 0$, $h > 0$, then*

$$P(X \geq t + h|X \geq t) = P(X \geq h).$$

Proof. For $t > 0$ we have

$$P(X \geq t) = \int_t^\infty \lambda e^{-\lambda x}dx = e^{-\lambda t}.$$

Hence for each $t > 0$ and each $h > 0$,

$$P(X \geq t + h|X \geq t) = \frac{P(X \geq t + h)}{P(X \geq t)} = \frac{e^{-\lambda(t+h)}}{e^{-\lambda t}} = e^{-\lambda h} = P(X \geq h).$$

□

What are the implications of the memoryless property? If human lifetimes were Exponential, then conditional on having survived to the age of 80, your remaining lifetime would have the same distribution as that of a newborn baby! Clearly, the memoryless property is not an appropriate description for human lifetimes.

The memoryless property is a very special property of the Exponential distribution. In fact, the Exponential is the only memoryless continuous distribution (with support $(0, \infty)$); and Geometric distribution is the only memoryless discrete distribution (with support $0, 1, \dots$).

Example 5.6. We try to model the waiting time at a bus station. When any bus arrives, suppose the time until the next bus arrives is an Exponential random variable with mean 10 minutes. You arrive at the bus stop at a random time, not knowing how long ago the previous bus came. What is the distribution of your waiting time for the next bus? What is the average time that you have to wait? What if you know the previous bus left 10 minutes ago, does that change your expected waiting time?

Solution: Let X be the waiting time and we know it is an Exponential distribution. Since $E(X) = 1/\lambda = 10$, the parameter $\lambda = 1/10$. Thus $X \sim \text{Expo}(0.1)$. By the memoryless property, how much longer the next bus will take to arrive is independent of how long ago the previous bus arrived. The average time you have to wait is always 10 minutes.

5.7 Poisson process

Now we point out the connection between the Poisson process and the exponential distribution — Let $X_1, X_2, \ldots$ be a sequence of events randomly occurred over time (random arrivals). If the number of events occurred in a given period of time follows a Poisson distribution, then the time interval between two events follows an Exponential distribution, *vice versa*.

Suppose the number of events occurred in an interval t is subject to Poisson distribution: $N \sim \text{Pois}(\lambda t)$. Let T be the waiting time before any event occurs. The waiting time being t is equivalent to $N = 0$ for time period t:

$$P(T > t) = P(N_t = 0) = e^{-\lambda t}\frac{(\lambda t)^0}{0!} = e^{-\lambda t}$$

where $N_t = \#$ emails in $[0, t]$. The CDF of T is

$$F(t) = 1 - P(T > t) = 1 - e^{-\lambda t}.$$

The PDF of T is

$$f(t) = F'(t) = \lambda e^{-\lambda t}.$$

This indicates $T \sim \mathit{Expo}(\lambda)$.

Definition 5.11. A sequence of arrivals in continuous time is a **Poisson process** with rate λ if

- The number of arrivals in an interval of length t is distributed $Pois(\lambda t)$;
- The numbers of arrivals in disjoint time intervals are independent.

Thus, Poisson distribution is used to model the number of random events in a period of time. Exponential distribution is used to model the time interval between two of these events.

When we introduced Poisson distribution in Chapter 3, we have said that Poisson distribution is used to model the scenario where the number of events is large and the probability of each event occurring is small. What is the connection here? The events occur randomly. Image we divide the time line into infinitely small interval (e.g. milliseconds), then an event either happens in a millisecond or not. Thus, we have a large number of Bernoulli trials. The total number of events occurred is approximated by a Binomial distribution, where n is huge, and p the probability that an event occurs in a particular millisecond is very small. This is the typical case of Poisson distribution.

Example 5.7. Suppose the number of calls to a phone number is a Poisson process with parameter λ. $\tau \sim Exp(\mu)$ is the duration of each call. It is reasonable to assume that τ is independent of the Poisson process. What is the probability that the $(n+1)$-th call gets a busy signal, i.e. it comes when the user is still responding to the n-th call?

Solution: Let T_n be the arrival time of the n-th customer. The probability we want to find is

$$P(T_n + \tau > T_{n+1}|\tau) = P(T_{n+1} - T_n < \tau|\tau) = P(X_n < \tau|\tau).$$

As we have discussed, X_n follows an Exponential distribution. Thus,

$$P(X_n < \tau|\tau) = 1 - e^{-\lambda\tau}.$$

To find the unconditional probability,

$$\begin{aligned} P(X_n < \tau) &= \int_0^\infty P(X_n < \tau|\tau) f(\tau) d\tau = \int_0^\infty (1 - e^{-\lambda\tau}) \mu e^{-\mu\tau} d\tau \\ &= 1 - \mu \int_0^\infty e^{-(\lambda+\mu)\tau} d\tau = \frac{\lambda}{\lambda+\mu}. \end{aligned}$$

5.8 Gamma distribution

The Gamma distribution is a continuous distribution on the positive real line; it is a generalization of the Exponential distribution. While an Exponential RV represents the waiting time for the first event to occur, we shall see that a Gamma RV represents the total waiting time for n events to occur.

Let's start with a simple case. Suppose we want to find out the total waiting until the 2nd event occurred. Let $Y = X_1 + X_2$ where $X_1, X_2 \sim Expo(\lambda)$ independently. If Y is discrete, we have $P(Y = y) = \sum_{k=0}^{y} P(X_1 = k, X_2 = y - k)$. For continuous y, we have

$$\begin{aligned} f_Y(y) &= \int_0^y f_X(x) f_X(y-x) dx = \int_0^y \lambda e^{-\lambda x} \lambda e^{-\lambda(y-x)} dx \\ &= \int_0^y \lambda^2 e^{-\lambda y} dx = \lambda^2 e^{-\lambda y} y. \end{aligned}$$

If there is a third variable,

$$\begin{aligned} f_Z(z) &= \int_0^z f_X(x) f_Y(z-x) dx = \int_0^z \lambda e^{-\lambda x} \lambda^2 e^{-\lambda(z-x)} (z-x) dx \\ &= \lambda^3 e^{-\lambda z} \int_0^z (z-x) dx = \lambda^3 e^{-\lambda z} z^2/2. \end{aligned}$$

The general pattern is the Gamma distribution.

Definition 5.12. An random variable X is said to have the **Gamma distribution** with parameters a and λ, $a > 0$ and $\lambda > 0$, if it has the PDF

$$f(x) = \frac{\lambda^a}{\Gamma(a)} x^{a-1} e^{-\lambda x}, \quad x > 0$$

We write $X \sim \text{Gamma}(a, \lambda)$.

Verify this is a valid PDF:

$$\int_0^\infty \frac{1}{\Gamma(a)} (\lambda x)^a e^{-\lambda x} \frac{dx}{x} \overset{u=\lambda x}{=} \frac{1}{\Gamma(a)} \int_0^\infty u^a e^{-u} \frac{du}{u} = \frac{\Gamma(a)}{\Gamma(a)} = 1.$$

Taking $a = 1$, the $\text{Gamma}(1, \lambda)$ PDF is $f(x) = \lambda e^{-\lambda x}$, which is the same as $\text{Expo}(\lambda)$. So Exponential distribution is a special case of Gamma distribution.

Let's find the expectation and variance of the Gamma distribution. Let $Y \sim$ Gamma$(a, 1)$. Recall Γ function has the property $\Gamma(a+1) = a\Gamma(a)$.

$$E(Y) = \int_0^\infty y \cdot \frac{1}{\Gamma(a)} y^{a-1} e^{-y} dy = \frac{1}{\Gamma(a)} \int_0^\infty y^a e^{-y} dy = \frac{\Gamma(a+1)}{\Gamma(a)} = a.$$

Apply LOTUS to evaluate the second moment:

$$E(Y^2) = \int_0^\infty y^2 \cdot \frac{1}{\Gamma(a)} y^{a-1} e^{-y} dy = \frac{1}{\Gamma(a)} \int_0^\infty y^{a+1} e^{-y} dy = \frac{\Gamma(a+2)}{\Gamma(a)} = (a+1)a.$$

Therefore,

$$Var(Y) = (a+1)a - a^2 = a.$$

So for $Y \sim$ Gamma$(a, 1)$, $E(Y) = Var(Y) = a$. For the general case $X \sim$ Gamma(a, λ), we now show that $X = \frac{Y}{\lambda}$. Note that

$$\begin{aligned} F_X(x) &= P(X \le x) = P(Y \le x/\lambda) = F_Y(x/\lambda) \\ f_X(x) &= \frac{dF_X}{dx} = \frac{\partial F_Y}{\partial y} \frac{dy}{dx} = f_Y(y)\lambda \end{aligned}$$

Therefore,

$$f_X(x) = \frac{1}{\Gamma(a)} y^{a-1} e^{-y} \lambda = \frac{\lambda^a}{\Gamma(a)} x^{a-1} e^{-\lambda x}.$$

Hence, we have $E(X) = \frac{a}{\lambda}$, $Var(X) = \frac{a}{\lambda^2}$.

Theorem 5.7. *Let $X_1, \ldots, X_n$ be independent and identical Expo(λ). Then*

$$X_1 + \cdots + X_n \sim \textit{Gamma}(n, \lambda).$$

Proof. Let's prove by showing the MGFs are equivalent.

$$M_X(t) = E(e^{tX}) = \int_0^\infty e^{tx} \lambda e^{-\lambda x} dx = \frac{\lambda}{\lambda - t} \quad \text{for } t < \lambda$$

Thus, the MGF of $Y = X_1 + \cdots + X_n$ is $M_Y(t) = (M_X(t))^n = \left(\frac{\lambda}{\lambda - t}\right)^n$. We verify this is the MGF of a Gamma distribution. Suppose $Y \sim$ Gamma(n, λ), it has MGF:

$$\begin{aligned}
M_Y(t) = E(e^{tY}) &= \int_0^\infty e^{ty} \frac{\lambda^n}{\Gamma(a)} y^{n-1} e^{-\lambda y} dy \\
&= \frac{\lambda^n}{(\lambda - t)^n} \int_0^\infty \frac{1}{\Gamma(a)} ((\lambda - t)y)^{n-1} e^{-(\lambda - t)y} (\lambda - t) dy \\
&= \frac{\lambda^n}{(\lambda - t)^n} \int_0^\infty \frac{1}{\Gamma(a)} u^{n-1} e^{-u} du \qquad u = (\lambda - t)y \\
&= \left(\frac{\lambda}{\lambda - t} \right)^n .
\end{aligned}$$

□

Thus, if X_i represents the $i.i.d$ inter-arrival time. Y has the interpretation of the arrival time until the n-th event.

$$Y = \sum_{i=1}^n X_i = \sum_{i=1}^n (\text{time of the i-th arrival}) \sim \text{Gamma}(n, \lambda).$$

Example 5.8 (Service time in a queue)**.** Customer i must wait time X_i for service once reaching the head of the queue. The average service rate is 1 customer per 10 minutes. Assume the service for each customer is independent. If you are the 5th in the queue. What is the expected waiting to be served?

Solution: $X_i \sim$ Expo(0.1). Then $E(X_i) = 10$. Let Y be the time until you are served. Then $Y \sim$ Gamma(5, 0.1). Thus, $E(Y) = \frac{5}{0.1} = 50$ minutes. The probabilities of some selected values:

$$P(Y \leq t) = \begin{cases} 5\% & t = 20 \\ 18\% & t = 30 \\ 71\% & t = 60 \end{cases} .$$

5.9 Beta distribution*

The Beta distribution is a continuous distribution on the interval $(0, 1)$. It is a generalization of the Unif$(0, 1)$ distribution, allowing the PDF to be non-constant on $(0, 1)$.

Definition 5.13. A random variable X is said to have the **Beta distribution** with parameters a and b, $a > 0$ and $b > 0$, if its PDF is

$$f(x) = \frac{1}{\beta(a,b)} x^{a-1}(1-x)^{b-1}, \quad 0 < x < 1$$

where the constant $\beta(a,b)$ is chosen to make the PDF integrate to 1. We write this as $X \sim \text{Beta}(a,b)$.

The Beta distribution takes different shapes for different a and b values. Here are some general patterns:

- If $a = b = 1$, the Beta$(1,1)$ PDF is constant on $(0,1)$, equivalent to Unif$(0,1)$.
- If $a < 1$ and $b < 1$, the PDF is U-shaped and opens upward. If $a > 1$ and $b > 1$, the PDF opens downward.
- If $a = b$, the PDF is symmetric about $1/2$. If $a > b$, the PDF favors values larger than $1/2$. If $a < b$, the PDF favors values smaller than $1/2$.

To make the PDF integrates to 1, the constant $\beta(a,b)$ has to satisfy

$$\beta(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}dx.$$

We now try to find this integral:

$$\begin{aligned}
\beta(a,b) &= \int_0^1 \underbrace{x^{a-1}}_{f}\underbrace{(1-x)^{b-1}}_{g'}dx \\
&= \left[-x^{a-1}\frac{(1-x)^b}{b}\right]_0^1 + \int_0^1 (a-1)x^{a-2}\frac{(1-x)^b}{b}dx \\
&= \frac{a-1}{b}\beta(a-1,b+1) \\
&= \frac{a-1}{b}\cdot\frac{a-2}{b+1}\beta(a-2,b+2) \\
&= \frac{a-1}{b}\cdot\frac{a-2}{b+1}\cdot\frac{a-3}{b+2}\beta(a-3,b+3) \\
&\vdots \\
&= \frac{(a-1)!}{b(b+1)(b+2)\cdots(b+a-2)}\underbrace{\beta(1,a+b-1)}_{\frac{1}{a+b-1}} \\
&= \frac{(a-1)!}{\frac{(b+a-2)!}{(b-1)!}}\cdot\frac{1}{a+b-1} \\
&= \frac{(a-1)!(b-1)!}{(a+b-1)!} \\
&= \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.
\end{aligned}$$

Example 5.9. Let $X_1,\dots,X_n$ be independent random variables with the uniform distribution on the interval $[0,1]$. Find the distribution of $Y=\max(X_1,\dots,X_n)$.

Solution: Let's find the CDF of Y:

$$\begin{aligned}
P(Y\le y) &= P(X_1\le y\cap X_2\le y\cap\cdots\cap X_n\le y) \\
&\overset{iid}{=} P(X_1\le y)P(X_2\le y)\cdots P(X_n\le y) \\
&= y^n
\end{aligned}$$

for $y\in[0,1]$. Hence,

$$F_Y(y)=P(Y\le y)=\begin{cases}0 & y<0\\ y^n & 0\le y\le 1\\ 1 & y>1\end{cases}$$

The PDF of Y is

$$f_Y(y) = F'_Y(y) = \begin{cases} ny^{n-1} & 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Thus, $Y \sim Beta(n, 1)$.

Beta distributions are often used as *priors* for parameters in Bayesian inference. We do not cover Bayesian inference in this book. Nonetheless we illustrate this with an example.

Example 5.10 (Beta-Binomial conjugacy)**.** We have a coin that lands Heads with probability p, but we don't know what p is. Our goal is to infer the value of p after observing the outcomes of n tosses of the coin. The larger that n is, the more accurately we should be able to estimate p.

Solution: We model the unknown parameter p as a Beta distribution, $p \sim \text{Beta}(a, b)$. Since we are completely ignorant about this p, we can also model it as the uniform distribution. But we will see that using the Beta distribution is even simpler than the uniform distribution. Let X be the number of heads in n tosses of the coin. Then

$$X|p \sim \text{Bin}(n, p)$$

Apply the Bayes' rule to inverse the conditioning:

$$\begin{aligned} f(p|X = k) &= \frac{P(X = k|p)f(p)}{P(X = k)} \\ &= \frac{\binom{n}{k} p^k (1-p)^{n-k} \cdot \frac{1}{\beta(a,b)} p^{a-1}(1-p)^{b-1}}{\int_0^1 \binom{n}{k} p^k (1-p)^{n-k} f(p) dp} \\ &\propto p^{a+k-1}(1-p)^{b+n-k-1} \end{aligned}$$

This the kernel of $\text{Beta}(a+k, b+n-k)$. The rest is just a normalizing constant. Therefore,

$$p|X = k \sim \text{Beta}(a + k, b + n - k).$$

The *posterior* distribution of p after observing $X = k$ is still a Beta distribution! This is a special relationship between the Beta and Binomial distributions called *conjugacy*: if we have a Beta prior distribution on p and data that are conditionally Binomial given p, then when going from prior to posterior, we don't leave the family of Beta distributions. We say that the *Beta is the conjugate prior of the Binomial*.

Chapter 6

Joint Distributions

6.1 Joint, marginal and conditional distributions

A joint distribution is a statistical concept used to describe the likelihood of two or more random variables occurring together. When we talk about joint distribution, we are considering the probability of different values of these variables happening simultaneously, rather than in isolation. Suppose we toss a coin and roll a die. Joint probability represents the probability that the two events happening simultaneously, e.g. $P(\text{Coin} = H, \text{Die} = 6)$.

Given the joint distribution, we are interested in: (i) the distribution of multi-variables simultaneously (joint probability); (ii) the distribution of one variable ignoring other variables (marginal probability); (iii) the distribution of one variable given the value of other variables (conditional probability).

	Discrete	Continuous
Joint CDF	$F_{XY}(x,y) = P(X \leq x, Y \leq y)$	$F_{XY}(x,y) = P(X \leq x, Y \leq y)$
Joint PMF/PDF	$p_{XY}(x,y) = P(X = x, Y = y)$	$f_{XY}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{XY}(x,y)$
	$\sum_x \sum_y P(X = x, Y = y) = 1$	$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f_{XY}(x,y) dx dy = 1$
Marginal PMF/PDF	$P(X = x) = \sum_y P(X = x, Y = y)$	$f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x,y) dy$
Conditional PMF/PDF	$P(X = x \vert Y = y) = \frac{P(X=x, Y=y)}{P(Y=y)}$	$f_{X\vert Y}(x\vert y) = \frac{f_{XY}(x,y)}{f_Y(y)}$
Independence	$P(X = x, Y = y) = P(X = x)P(Y = y)$	$f_{XY}(x,y) = f_X(x) f_Y(y)$
	$P(X = x \vert Y = y) = P(X = x)$	$f_{X\vert Y}(x\vert y) = f_X(x)$
	$F_{XY}(x,y) = F_X(x) F_Y(y)$	$F_{XY}(x,y) = F_X(x) F_Y(y)$
Bayes' rule	$P(Y = y \vert X = x) = \frac{P(X=x\vert Y=y)P(Y=y)}{P(X=x)}$	$f_{Y\vert X}(y\vert x) = \frac{f_{X\vert Y}(x\vert y) f_Y(y)}{f_X(x)}$
LOTP	$P(X = x) = \sum_y P(X = x \vert Y = y) P(Y = y)$	$f_X(x) = \int_{-\infty}^{+\infty} f_{X\vert Y}(x\vert y) f_Y(y) dy$
LOTUS	$E(g(X,Y)) = \sum_x \sum_y g(x,y) P(X = x) P(y = y)$	$E(g(X,Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{XY}(x,y) dx dy$

Table 6.1: Joint, marginal and conditional distributions

Example 6.1. Let X be an indicator of an individual being a current smoker. Let Y be the indicator of his developing lung cancer at some point in his life. The joint PMF of X and Y is as specified in the table below.

	$Y=1$	$Y=0$	**Total**
$X=1$	0.05	0.20	**0.25**
$X=0$	0.03	0.72	**0.75**
Total	**0.08**	**0.92**	**1**

The marginal PMF for having lung cancer is

$$P(Y=1) = P(Y=1,X=0)+P(Y=1,X=1)=0.08,$$
$$P(Y=0) = P(Y=0,X=0)+P(Y=0,X=1)=0.92.$$

The conditional PMF of having lung cancer conditioned on being a smoker is

$$P(Y=1|X=1)=\frac{P(X=x,Y=y)}{P(X=x)}=\frac{0.05}{0.25}=\frac{1}{5}.$$

In this example, X, Y are not independent, because

$$P(X=1,Y=1)\neq P(X=1)P(Y=1).$$

Example 6.2. Suppose X and Y are uniformly distributed on a disk $\{(x,y): x^2+y^2\leq 1\}$. Find the joint PDF, marginal distributions and conditional distributions. Are X and Y independent?

Solution: The area of the disk is π, therefore

$$f(x,y)=\begin{cases}\frac{1}{\pi} & x^2+y^2\leq 1\\ 0 & \text{otherwise}\end{cases}$$

The marginal distributions are

$$\begin{aligned} f_X(x) &= \int_{-\sqrt{1-x^2}}^{\sqrt{1+x^2}} \frac{1}{\pi} dy = \frac{2}{\pi}\sqrt{1-x^2}, \qquad -1 \leq x \leq 1 \\ f_Y(y) &= \int_{-\sqrt{1-y^2}}^{\sqrt{1+y^2}} \frac{1}{\pi} dx = \frac{2}{\pi}\sqrt{1-y^2}, \qquad -1 \leq y \leq 1 \end{aligned}$$

The conditional distributions are

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{\frac{1}{\pi}}{\frac{2}{\pi}\sqrt{1-x^2}} = \frac{1}{2\sqrt{1-x^2}}$$

Therefore, $Y|X \sim \text{Unif}(-\sqrt{1-x^2}, \sqrt{1-x^2})$.

Since $f(x,y) \neq f_X(x)f_Y(y)$, X and Y are not independent. This is because knowing the value of X constrains the value of Y.

Example 6.3. Suppose $X, Y \overset{iid}{\sim} Unif(0,1)$. Find the probability $P\left(Y \leq \frac{1}{2X}\right)$.

Solution: The joint distribution is

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

$$P\left(Y \leq \frac{1}{2X}\right) = \int_0^{1/2}\int_0^1 1dydx + \int_{1/2}^1\int_0^{1/2x} 1dydx = \frac{1}{2} + \int_{1/2}^1 \frac{1}{2x}dx = \frac{1}{2} + \ln\sqrt{2}.$$

Example 6.4. For $X, Y \overset{iid}{\sim} \text{Unif}(0,1)$, find $E(|X-Y|)$.

Solution: Apply 2D LOTUS:

$$\begin{aligned} E(|X-Y|) =& \int_0^1\int_0^1 |x-y|dxdy \\ =& \int_0^1\int_y^1 (x-y)dxdy + \int_0^1\int_0^y (y-x)dxdy \\ =& 2\int_0^1\int_y^1 (x-y)dxdy \\ =& \frac{1}{3}. \end{aligned}$$

Example 6.5. $X, Y \overset{iid}{\sim} N(0,1)$, find $E(|X-Y|)$.

Solution: Since the sum or difference of independent Normals is Normal, $X - Y \sim N(0,2)$. Let $Z = X - Y$. Then $Z \sim N(0,1)$, and $E(|X-Y|) = \sqrt{2}E(|Z|)$. Apply LOTUS,

$$E(|Z|) = \int_{-\infty}^{\infty} |z| \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz = 2\int_{0}^{\infty} z \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz = \sqrt{\frac{2}{\pi}},$$

Therefore, $\mathbb{E}(|X-Y|) = \frac{2}{\sqrt{\pi}}$.

6.2 Joint normal distribution

Definition 6.1. (X, Y) is said to have a **Bivariate Normal** distribution if the joint PDF satisfies

$$f(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}(x^2 + y^2 - 2\rho xy)\right)$$

where $\rho \in (-1,1)$ is the correlation between X and Y.

A **Multivariate Normal (MVN)** is fully specified by knowing the mean of each component, the variance of each component, and the covariance or correlation between any two components. In other words, the parameters of an MVN random vector $(X_1, ..., X_k)$ are as follows:

- the mean vector $(\mu_1, ..., \mu_k)$, where $E(X_j) = \mu_j$;
- the covariance matrix $Cov(X_i, X_j)$ for $1 \leq i, j \leq k$.

If $(X_1, ..., X_k)$ is MVN, then the marginal distribution of every X_j is Normal. However, the converse is false: it is possible to have Normally distributed $X_1, ..., X_k$ such that $(X_1, ..., X_k)$ is not Multivariate Normal.

Theorem 6.1. *A random vector $(X_1, ..., X_k)$ is Multivariate Normal if every linear combination of the X_j has a Normal distribution. That is, we require $t_1X_1 + \cdots + t_kX_k$ to have a Normal distribution for any choice of constants $t_1, ..., t_k$.*

Theorem 6.2. *Within an MVN random vector, uncorrelated implies independent. In particular, if (X, Y) is Bivariate Normal and $Corr(X, Y) = 0$, then X and Y are independent.*

This is a special property of MVN random variables. In general, uncorrelated does not imply independent.

Theorem 6.3. *If (X, Y) is Bivariate Normal, then the conditional expectation satisfies*

$$E(Y|X) = E(Y) + \frac{Cov(X, Y)}{Var(X)}(X - E(X)).$$

This is also a special property of MVN —$E(Y|X)$ is a linear function of X. This is not the case in general.

6.3 Conditional expectation

Theorem 6.4. *For any random variable X and Y,*

$$E(E(Y|X)) = E(Y).$$

This is known as the ***law of iterated expectation****.*

Proof. Note that $E(Y|X) = g(X)$ is a function of X. Apply LOTUS:

$$\begin{aligned} E(E(Y|X)) &= \int g(x)f(x)dx \\ &= \int \left(\int yf(y|x)dy \right) f(x)dx \\ &= \int \int yf(y|x)f(x)dydx \\ &= \int y \int f(y,x)dx\,dy \\ &= \int_{-\infty}^{\infty} yf(y)dy \\ &= E(Y). \end{aligned}$$

$\square$

Theorem 6.5. *For any random variable X and Y, and any function g,*

$$E(g(X)Y|X) = g(X)E(Y|X).$$

Proof. For any specific value of $X = x$, $g(x)$ is a constant. Thus, $E(g(x)Y|X = x) = g(x)E(Y|X = x)$. This is true for all values of x. □

Example 6.6 (PG exam)**.** Suppose $X \sim \text{Unif}(0, 1)$, and $Y|X \sim N(X, X^2)$, meaning that for a given $X = x$, Y is normally distributed with mean x and variance x^2. Find $E(Y)$, $Var(Y)$ and $Cov(X, Y)$.

Solution:

Since $Y|X \sim N(X, X^2)$, we know $E(Y|X) = X$. By the law of iterated expectation,

$$E(Y) = E(E(Y|X)) = E(X) = \frac{1}{2}.$$

For the variance,

$$Var(Y) = E(Y^2) - (E(Y))^2 = E(E(Y^2|X)) - \frac{1}{4}.$$

Since

$$Var(Y|X) = E(Y^2|X) - E^2(Y|X) = E(Y^2|X) - X^2 = X^2,$$

we have $E(Y^2|X) = 2X^2$. Meanwhile, $E(X^2) = \int_0^1 x^2 \cdot 1dx = \frac{1}{3}$. Therefore,

$$Var(Y) = \frac{2}{3} - \frac{1}{4} = \frac{5}{12}.$$

For the covariance,

$$E(XY) = E(E(XY|X)) = E(XE(Y|X)) = E(X^2) = \frac{1}{3},$$

$$Cov(X, Y) = E(XY) - E(X)E(Y) = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}.$$

Theorem 6.6. *Conditional expectation $E(Y|X)$ is the best predictor for Y using X (minimized the square loss function).*

Proof. Let $g(X)$ be a predictor for Y using X. We want to find the g such that minimizes $E(Y - g(X))^2$.

$$\begin{aligned}
E(Y - g(X))^2 &= E(Y - E(Y|X) + E(Y|X) - g(X))^2 \\
&= E(Y - E(Y|X))^2 + 2\underbrace{E(Y - E(Y|X)}_{E(Y)=E(E(Y|X))}((E(Y|X) - g(X)) + E(E(Y|X) - g(X))^2 \\
&= E(Y - E(Y|X))^2 + E(E(Y|X) - g(X))^2 \\
&\geq E(Y - E(Y|X))^2.
\end{aligned}$$

Therefore, $E(Y - g(X))^2$ is minimized when $g(X) = E(Y|X)$. □

6.4 Linear conditional expectation model

Definition 6.2. An extremely widely used method for data analysis in statistics is linear regression. In its most basic form, we want to predict the mean of Y using a single explanatory variable X. **A linear conditional expectation model** assumes that $E(Y|X)$ is linear in X:

$$E(Y|X) = a + bX,$$

or equivalently,

$$Y = a + bX + \epsilon,$$

with $E(\epsilon|X) = 0$. The intercept and the slope is given by

$$b = \frac{Cov(X, Y)}{Var(X)}, a = E(Y) - bE(X).$$

We first show the equivalence of the two expressions of the model. Let $Y = a + bX + \epsilon$, with $E(\epsilon|X) = 0$. Then by linearity,

$$E(Y|X) = E(a|X) + E(bX|X) + E(\epsilon|X) = a + bX.$$

Conversely, suppose that $E(Y|X) = a + bX$, and define

$$\epsilon = Y - (a + bX).$$

Then $Y = a + bX + \epsilon$, with

$$E(\epsilon|X) = E(Y|X) - E(a + bX|X) = E(Y|X) - (a + bX) = 0.$$

To derive the expression for a and b, take covariance between X and Y,

$$\begin{aligned} Cov(X, Y) &= Cov(X, a + bX + \epsilon) \\ &= Cov(X, a) + bCov(X, X) + Cov(X, \epsilon) \\ &= bVar(X) + Cov(X, \epsilon) \end{aligned}$$

Note that $Cov(X, \epsilon) = 0$ because

$$\begin{aligned} Cov(X, \epsilon) &= E(X\epsilon) - E(X)E(\epsilon) \\ &= E(E(X\epsilon|X)) - E(X)E(E(\epsilon|X)) \\ &= E(XE(\epsilon|X)) - E(X)E(E(\epsilon|X)) \\ &= 0 \end{aligned}$$

Therefore,

$$Cov(X, Y) = bVar(X)$$

Thus,

$$\begin{aligned} b &= \frac{Cov(X, Y)}{Var(X)}, \\ a &= E(Y) - bE(X) = E(Y) - \frac{Cov(X, Y)}{Var(X)}E(X). \end{aligned}$$

6.5 Change of variables

Theorem 6.7. *Let X be a continuous r.v. with PDF f_X, and let $Y = g(X)$, where g is differentiable and strictly increasing (or strictly decreasing). Then the PDF of Y is given by*

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

where $x = g^{-1}(y)$.

Proof. Let g be strictly increasing. The CDF of Y is

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

By the chain rule, the PDF of Y is

$$f_Y(y) = f_X(x)\frac{dx}{dy}.$$

If g is strictly decreasing,

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = P(X \geq g^{-1}(y)) = 1-F_X(g^{-1}(y)) = 1-F_X(x)$$

Then the PDF of Y is

$$f_Y(y) = -f_X(x)\frac{dx}{dy}.$$

But in this case, $dx/dy < 0$. So taking absolute value covers both cases. □

Example 6.7 (Log-Normal PDF)**.** Let $X \sim N(0,1)$, $Y = e^X$. Then the distribution of Y is called the **Log-Normal distribution**. Find the PDF of Y.

Since $g(x) = e^x$ is strictly increasing. Let $y = e^x$, so $x = \log y$ and $dy/dx = e^x$. Then

$$f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right| = \varphi(x)\frac{1}{e^x} = \varphi(\log y)\frac{1}{y}, \quad y > 0.$$

Note that after applying the change of variables formula, we write everything on the right-hand side in terms of y, and we specify the support of the distribution. To determine the support, we just observe that as x ranges from $-\infty$ to∞, e^x ranges from 0 to ∞.

Example 6.8 (Chi-Square PDF)**.** Let $X \sim N(0,1)$, $Y = X^2$. The distribution of Y is an example of a **Chi-Square distribution**. Find the PDF of Y.

In this case, we can no longer apply the change of variables formula because $g(x) = x^2$ is not one-to-one. Instead, we use the CDF:

$$F_Y(y) = P(X^2 \leq y) = P(-\sqrt{y} \leq X \leq \sqrt{y}) = \Phi(\sqrt{y}) - \Phi(-\sqrt{y}) = 2\Phi(\sqrt{y}) - 1$$

Therefore,

$$f_Y(y) = 2\varphi(\sqrt{y}) \cdot \frac{1}{2}y^{-1/2} = \varphi(\sqrt{y})y^{-1/2}, \quad y > 0.$$

Theorem 6.8. *Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a continuous random vector with joint PDF $f_{\mathbf{X}}$, and let $\mathbf{Y} = g(\mathbf{X})$ where g is an invertible function from $\mathbb{R}^n$ to $\mathbb{R}^n$. Let $\mathbf{y} = g(\mathbf{x})$. Define the Jacobian matrix:*

$$\frac{\partial \mathbf{x}}{\partial \mathbf{y}} = \begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial x_n}{\partial y_1} & \frac{\partial x_n}{\partial y_2} & \cdots & \frac{\partial x_n}{\partial y_n} \end{pmatrix}.$$

Also assume that the determinant of the Jacobian matrix is never 0. Then the joint PDF of $\mathbf{Y}$ is

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}(\mathbf{x}) \left| \frac{\partial \mathbf{x}}{\partial \mathbf{y}} \right|,$$

where $\left| \frac{\partial \mathbf{x}}{\partial \mathbf{y}} \right|$ is the absolute value of the determinant of the Jacobian matrix.

Example 6.9. Suppose $X, Y \overset{iid}{\sim} Expo(1)$. Find the distribution of $X/(X+Y)$.

Solution: Let $U = \frac{X}{X+Y}$, $V = X + Y$. Then $X = UV$, $Y = V - UV$. The determinant of the Jacobian matrix is

$$\left| \frac{\partial(x, y)}{\partial(u, v)} \right| = \begin{vmatrix} v & u \\ -v & 1-u \end{vmatrix} = v$$

Thus, the joint distribution of (U, V) is

$$f_{UV}(u, v) = f_{XY}(x, y)|v| = f_X(x) f_Y(y) v = e^{-(x+y)} v = e^{-v} v.$$

The distribution of $X/(X+Y)$ is equivalent to the marginal distribution of U:

$$f_U(u) = \int_0^\infty e^{-v} v dv = 1$$

for $0 \le u \le 1$. Hence U is a Uniform distribution over [0,1].

Chapter 7

Sampling distribution

7.1 Samples and statistics

We model real-world uncertain events with random variables. We have also introduced various distributions suitable to model different kinds of events. However, we never observe the full distribution or the true parameters of the assumed distribution. Instead, we only observe a sample of that random variable. We can only infer the properties of the distribution from a limited sample. For example, suppose we model the hight of an Asian women with a normal distribution. But we never know exactly what the mean and variance are. We can only observe a sample of the distribution.

In statistics, the conceptual distribution F is called the **population distribution**, or just the **population**.[1] It is tempting to think of the population as all the observations (e.g. all the population on the planet), but this is not exactly correct. The population distribution is more of a mathematical abstraction or an assumption. Suppose we are modeling the height of human being, even if we have all the observations on the planet, that does not include the people that have died or yet to be born. Thus, it is still a sample of the assumed distribution.

A collection of random variables $\{X_1, X_2, \dots, X_n\}$ is a **random sample** from the population F if X_i are **independent and identically distributed ($\boldsymbol{i.i.d}$)**

[1] This section is based on Bruce Hansen's *Probability and Statistics for Economists.*

with distribution F. What we mean by *i.i.d* is that $X_1, \dots, X_n$ are mutually independent and have exactly the same distribution $X_i \sim F$. Survey sampling is an useful metaphor to understand random sampling, in which we randomly select a subset of the population with equal probability. The **sample size** n is the number of individuals in the sample.

A **data set** is a collection of numbers, typically organized by observation. We sometimes call a data set also as a sample. But it should not be confused with the random sample defined above. As the former is a collection of random variables, whereas the latter is one **realization** of the random variables.

Typically, we will use X without the subscript to denote a random variable or vector with distribution F, X_i with a subscript to denote a random observation in the sample, and x_i or x to denote a specific or realized value.

The problem of **statistical inference** is to learn about the underlying process — the population distribution or data generating process — by examining the observations. In most cases, we assume the population distribution and want to learn about the its parameters (e.g. μ and σ^2 in the normal distribution). As a convention, we use greek letters to denote population parameters.

A **statistic** is a function of the random sample $\{X_1, X_2, \dots, X_n\}$. Recall that there is a distinction between random variables and their realizations. Similarly there is a distinction between a statistic as a function of a random sample — and is therefore a random variable as well — and a statistic as a function of the realized sample, which is a realized value. When we treat a statistic as random we are viewing it is a function of a sample of random variables. When we treat it as a realized value we are viewing it as a function of a set of realized values. One way of viewing the distinction is to think of "before viewing the data" and "after viewing the data". When we think about a statistic "before viewing" we do not know what value it will take. From our vantage point it is unknown and random. After viewing the data and specifically after computing and viewing the statistic the latter is a specific number and is therefore a realization. It is what it is and it is not changing. The randomness is the process by which the data was generated — and the understanding that if this process were repeated the sample would be different and the specific realization would be therefore different. The distribution of a statistic is called the **sampling distribution**, since it is the distribution induced by sampling.

An **estimator** $\hat{\theta}$ for a population parameter θ is a statistic intended to infer θ.

It is conventional to use the hat notation $\hat{\theta}$ to denote an estimator. Note that $\hat{\theta}$ is a statistic and hence also a random variable. We call $\hat{\theta}$ an **estimate** when it is a specific value (or realized value) calculated in a specific sample.

A standard way to construct an estimator is by the analog principle. The idea is to express the parameter θ as a function of the population F, and then express the estimator $\hat{\theta}$ as the analog function in the sample.

For example, suppose we want to construct an estimator for the population mean $\mu = E(X)$. By definition, if each value of X is of equal probability, μ is simply the average. By analogy, we construct the **sample mean** as $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. It is conventional to denote a sample average by the notation "X bar". Because it is an estimator for μ, we also denote it as $\hat{\mu} = \bar{X}_n$. Note that from different samples we calculate different estimates. In one sample, $\hat{\mu} = 6.5$; in another sample, $\hat{\mu} = 6.7$. All of them are erroneous estimate of the true parameter μ. The question is therefore how close they are to the true parameter. To answer this question, we need to study the distribution of the sample mean.

7.2 Law of large numbers

We now introduce two important theorems describing the behavior of the sample mean as the sample size grows. Throughout this section and the next, we assume $X_1, X_2, \ldots, X_n$ are i.i.d RVs drawn from a population with mean μ and variance σ^2. The sample mean is defined as

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n}.$$

As we have discussed previously, the sample mean is itself a random variable with mean and variance:

$$\begin{aligned} E(\bar{X}_n) &= \frac{1}{n}E(X_1 + \cdots + X_n) = \frac{1}{n}(E(X_1) + \cdots + E(X_n)) = \mu, \\ Var(\bar{X}_n) &= \frac{1}{n^2}Var(X_1 + \cdots + X_n) \overset{iid}{=} \frac{1}{n^2}(Var(X_1) + \cdots + Var(X_n)) = \frac{\sigma^2}{n}. \end{aligned}$$

The law of large numbers (LLN) says that as n grows, the sample mean $\bar{X}_n$ converges to the true mean μ.

Theorem 7.1 (Strong law of large numbers)**.** *The sample mean $\bar{X}_n$ converges to the true mean μ point-wise as $n \to \infty$, with probability 1. In other words, the event $\bar{X}_n \to \mu$ has probability 1.*

Theorem 7.2 (Weak law of large numbers)**.** *For all $\epsilon > 0$, $P(|\bar{X}_n - \mu| > \epsilon) \to 0$ as $n \to \infty$. (this is known as converge in probability).*

We don't need a rigorous proof here. But an intuitive proof is obvious. As $n \to \infty$, $Var(\bar{X}_n) = \frac{\sigma^2}{n} \to 0$. The random variable $\bar{X}_n$ becomes fixed at μ as n becomes large. Thus, it converges to μ in a probabilistic sense.

It seems that the LLN just states the obvious. But it has wide applications in daily time that you might not even realize. What it says is essentially this: the uncertainty at the individual level becomes certain when aggregating together; the risks that are unmanageable at the individual level becomes manageable collectively. Think about a rare disease, it happens at 1 out of a million probability. For each individual, no one knows if they will get the disease or not. But as the sample size gets large, suppose we have one billion population, the LLN says the sample mean will be very close the true mean. That is, there will be almost surely 1000 people being infected by the disease. We provide two more examples.

Example 7.1 (Lottery)**.** A lottery company is designing a game with a 6-digit format. Each time someone buys a ticket, they receive a randomly generated 6-digit number. Only one number will win the grand prize of 10 million dollars. What should the company charge per ticket to break even?

Solution: The probability of winning the game is $p = 1/10^6$. Suppose the company has sold n tickets. The price for each ticket is x. The revenue for the company is therefore xn. By the LLN, the cost of the company should be very close to $10^7 np$. The break even point is $xn = 10^7 np$. So $x = 10^7 p = 10$. Therefore, if the company sells each ticket above 10 dollars. The business is surely profitable as long as n is large. If the company is a monopoly, it can reap as much profit as it desires as long as they know the basic probability theory! The same can be said about gambling companies.

Example 7.2 (Insurance)**.** Insurance is anther great application of the LLN. It is essentially the same as the the lottery game but most people do not realize it. Suppose there is a disease with infection rate of 1 out of 1 million. The medical expenditure to cure the disease is 10 million dollars. How much the insurance company should charge per customer to cover this disease?

Solution: The solution is essentially the same as above. Suppose the premium for the insurance product is x. The revenue of the company by selling the premium is xn. The cost is — when one customer is infected, the company has to pay the medical cost —$10^7 np$. The break even price for the insurance premium is thus 10 dollars.

What is the implication of this insurance? Without the insurance, each individual either chooses to set aside 10 million dollars pre-cautiously for the disease (if he is rich enough) or be exposed to the risk completely uncovered. The insurance product enables everyone to get covered at a cost of just 10 dollars. It is a typical example that the unmanageable risk at the individual level becomes manageable collectively.

7.3 Central limit theorem

The LLN shows the convergence of the sample mean to the population mean. What about the entire sample distribution? This is addressed by the central limit theorem (CLT), which, as its name suggests, is a limit theorem of **central importance** in statistics.

The CLT states that for large n, the distribution of $\bar{X}_n$ after standardization approaches a standard Normal distribution, regardless of the underlying distribution of X_i. By **standardization**, we mean that we subtract μ, the expected value of $\bar{X}_n$, and divide by $\sigma/\sqrt{n}$, the standard deviation of $\bar{X}_n$.

Theorem 7.3 (Central limit theorem)**.** *As $n \to \infty$,*

$$\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \to N(0,1) \text{ in distribution.}$$

In other words, the CDF of the left-hand side approaches the CDF of the standard normal distribution.

Proof. We will prove the CLT assuming the MGF of the X_i exists, though the theorem holds under much weaker conditions. Without loss of generality let

$\mu = 1, \sigma^2 = 1$ (since we standardize it anyway). We show that the MGF of $\sqrt{n}\bar{X}_n = (X_1 + \cdots + X_n)/\sqrt{n}$ converges to the MGF of the $N(0,1)$.

The MGF of $N(0,1)$ is

$$\begin{aligned}
E(e^{tX}) &= \int_{-\infty}^{\infty} e^{tx} \cdot \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2+tx} dx \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-t)^2+\frac{1}{2}t^2} dx \\
&= e^{\frac{t^2}{2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-t)^2} dx \\
&= e^{t^2/2}
\end{aligned}$$

Compute the MGF of $\sqrt{n}\bar{X}_n$:

$$\begin{aligned}
E(e^{\sqrt{n}\bar{X}_n}) &= E(e^{t(X_1+\cdots+X_n)/\sqrt{n}}) \\
&= E(e^{tX_1/\sqrt{n}})E(e^{tX_2/\sqrt{n}}) \cdots E(e^{tX_n/\sqrt{n}}) \\
&= \left[E(e^{tX_i/\sqrt{n}})\right]^n \qquad \text{since } i.i.d \\
&= \left[E\left(1 + \frac{tX_i}{\sqrt{n}} + \frac{t^2X_i^2}{2n} + o(n^{-1})\right)\right]^n \\
&= \left[1 + \frac{t}{\sqrt{n}}E(X_i) + \frac{t^2}{2n}E(X_i^2) + o(n^{-1})\right]^n \\
&= \left[1 + \frac{t^2}{2n} + o(n^{-1})\right]^n \\
&= \left[1 + \frac{t^2/2}{n} + o(n^{-1})\right]^n \\
&\to e^{t^2/2} \qquad \text{as } n \to \infty
\end{aligned}$$

Therefore, the MGF of $\sqrt{n}\bar{X}_n$ approaches the MGF of the standard normal. Since MGF determines the distribution, the distribution of $\sqrt{n}\bar{X}_n$ also approaches the standard normal distribution. □

The CLT tells us about the limiting distribution of $\bar{X}_n$ as $n \to \infty$. That means, we can reasonably approximate the distribution $\bar{X}_n$ with normal distribution when n is a finite large number —

$$\bar{X}_n \approx N(\mu, \sigma^2/n) \quad \text{for large } n.$$

The Central Limit Theorem was first proved by Pierre-Simon Laplace in 1810. Let's take a moment to admire the generality of this result. The distribution of the individual X_i can be anything in the world, as long as the mean and variance are finite. This does mean the distribution of X_i is irrelevant, however. If the distribution is fairly close to normal, the result would hold for smaller n. If the distribution is far away from normal, it would take larger n to converge.

The CLT gives the distribution of the sample mean regardless of the underlying distribution. This allows to assess the "quality" of the sample mean — how close it is to the true mean. The LLN tells us the larger the sample, the closer the sample mean to the population mean. The CLT tells us the distribution of the sample mean for sample size n. For smaller n, the distribution is more spread-out (a normal distribution with large σ^2); hence the uncertainty is huge, other values are more likely. For larger n, the uncertainty is reduced, most values would be centered around the true mean. We will delve deeper into this when we get to hypothesis testing.

Example 7.3. Suppose that a fair coin is tossed 900 times. Approximate the probability of obtaining more than 395 heads.

Solution: Let $H = \sum_{i=1}^{900} X_i$ be the number of heads, where $X_i \sim \text{Bern}(\frac{1}{2})$. We could compute the probability by

$$P(H > 495) = \sum_{k=496}^{900} \binom{900}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{900-k}$$

But this is quite tedious. Because $n = 900$ is reasonably large, we can apply the CLT:

$$\frac{1}{n}\sum_{i=1}^{900} X_i \sim N(\mu, \sigma^2/n) \quad \text{or}$$
$$\sum_{i=1}^{900} X_i \sim N(n\mu, n\sigma^2)$$

We know $\mu = E(X_i) = \frac{1}{2}$, $\sigma^2 = Var(X_i) = \frac{1}{4}$. Thus $H \sim N(450, 225)$. Therefore,

$$P(H > 495) = 1 - P(H \leq 495) \approx 1 - \Phi\left(\frac{495 - 450}{15}\right) = 0.0013.$$

7.4 Estimator accuracy

The central question of statistics is we want to learn about the population from a finite sample. We know sample mean is different from the population mean. But we also want to know how large the error could be, that is, how far or close the sample mean is from the true population mean. The question is exceedingly difficult to answer because the population mean is unknown. Fortunately, with the help of the CLT, we can say more about the distribution of the sample mean. This chapter bridges our probability theory with statistics. We use the theorems we have derived to infer the properties of a statistic.

As the purpose of statistics is to learn about the population, we want our sample estimator to be as good as possible. But what is a “good” estimator? This section we discuss two properties that we usually demand from a good estimator, namely, unbiasedness and consistency. Next section will tackle the more challenging concept of confidence interval.

Definition 7.1. The **bias** of an estimator $\hat{\theta}$ of a parameter θ is

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

We say that an estimator is **biased** if its sampling is incorrectly centered. We say that an estimator is **unbiased** is the bias is zero.

Theorem 7.4. *$\bar{X}_n$ is **unbiased** for $\mu = E(x)$ if $E(X) < \infty$.*

Proof.

$$E(\bar{X}_n) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu.$$

□

Theorem 7.5. *If $\hat{\theta}$ is an unbiased estimator of θ, then $\hat{\beta} = a\hat{\theta}+b$ is an unbiased estimator of $\beta = a\theta + b$.*

But obtaining an unbiased estimator is not always as straightforward as it seems. Consider the sample variance as an estimator for the population variance. By the analog principle, the sample variance should be

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu + \mu - \bar{X}_n)^2 \\
&= \frac{1}{n}\sum(X_i - \mu)^2 + \frac{2}{n}\sum(X_i - \mu)(\mu - \bar{X}_n) + \frac{1}{n}\sum(\mu - \bar{X}_n)^2 \\
&= \frac{1}{n}\sum(X_i - \mu)^2 + 2(\bar{X}_n - \mu)(\mu - \bar{X}_n) + (\mu - \bar{X}_n)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - (\bar{X}_n - \mu)^2 \\
&= \tilde{\sigma}^2 - (\bar{X}_n - \mu)^2
\end{aligned}$$

We know that

$$E(\tilde{\sigma}^2) = \frac{1}{n}\sum_{i=1}^{n} E(X_i - \mu)^2 = \sigma^2$$

Thus, if we compute the bias of this estimator:

$$\begin{aligned}
E[\hat{\sigma}^2] &= \sigma^2 - \frac{\sigma^2}{n} = \left(1 - \frac{1}{n}\right)\sigma^2 \\
\text{Bias}[\hat{\sigma}^2] &= -\frac{\sigma^2}{n} \neq 0
\end{aligned}$$

Therefore, the estimator $\hat{\sigma}^2$ is a biased estimator for σ^2! To correct the bias, we divide the sample sum of squares by $(n-1)$.

$$s^2 = \frac{n}{n-1}\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X_n})^2.$$

It is straightforward to see that s^2 is an unbiased estimator for σ^2. We call s^2 the **bias-corrected variance estimator**.

Theorem 7.6. *s^2 is an unbiased estimator for σ^2 if $E(X^2) < \infty$.*

Definition 7.2. The mean square error of an estimator $\hat{\theta}$ for θ is

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

By expanding the square we find that

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

Thus the MSE is the variance plus the squared bias. The MSE as a measure of accuracy combines the variance and bias.

Theorem 7.7. *For any estimator with a finite variance, we have*

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

Definition 7.3. An estimator is **consistent** if $\text{MSE}[\hat{\theta}] \to 0$ as $n \to \infty$.

Bias is the property of an estimator for finite samples. Consistency is the property of an estimator when the sample size gets large. It means that for any given data distribution, there is a sample size n sufficiently large such that the estimator $\hat{\theta}$ will be arbitrarily close to the true value θ with high probability. In practice, we usually do not know how large this n has to be. But it is a desirable property for an estimator to be considered a "good" estimator.

For unbiased estimator, MSE is solely determined by the variance of the estimator. Recall that the variance for the sample mean is $Var(\bar{X}_n) = \sigma^2/n$. But this is not a very useful formula because the it depends on unknown parameter σ^2. We need to replace these unknown parameters by estimators. To put the latter in the same units as the parameter estimate we typically take the square root before reporting. We thus arrive at the following concept.

Definition 7.4. A **standard error** of an estimator $\hat{\theta}$ is defined as

$$SE(\hat{\theta}) = \hat{V}^{1/2}$$

where $\hat{V}$ is the estimator for $Var[\hat{\theta}]$.

Definition 7.5. The **standard error** for $\bar{X}_n$ is

$$SE(\bar{X}_n) = \frac{s}{\sqrt{n}}$$

where s is the bias-corrected estimator for σ.

Note the difference between **standard error** and **standard deviation**. Standard deviation describes the dispersion of a distribution. Standard error is the standard deviation of an *estimator*. It indicates the "precision" of the estimator, thereby carrying a sense of "error". The smaller the standard error, the more precise the estimator.

7.5 Confidence intervals

Confidence intervals provide a method of adding more information to an estimator $\hat{\theta}$ when we wish to estimate an unknown parameter θ. We can find an interval (A, B) that we think has high probability of containing θ. The length of such an interval gives us an idea of how closely we can estimate θ.

Definition 7.6. A $100(1-\alpha)\%$ **confidence interval (CI)** for θ is an interval $[L(\theta), U(\theta)]$ such that the probability that the interval contains the true θ is $(1-\alpha)$.

Due to randomness we rarely seek a confidence interval with 100% coverage as this would typically need to be the entire parameter space. Instead we seek an interval which includes the true value with reasonably high probability. Standard choices are $\alpha = 0.05$ and 0.10, corresponding to 95% and 90% confidence.

Confidence intervals are reported to indicate the degree of precision of our estimates. The narrower the confidence interval, the more precise the estimate. Because a small range of values contains the true parameter with high probability.

With the help of the CLT, it is not hard to find the CI for the sample mean $\bar{X}_n$. Let's set $\alpha = 5\%$, that is, we are trying to find the CI that contains the true mean 95% of the times. Assume our sample size n is large enough to invoke the CLT, we thus have

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

Let's find the interval $[a, b]$ such that

$$P\left(a \leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq b\right) = 1 - 2\Phi(L) = 0.95$$

since the normal distribution is symmetric, $b = -a$. By looking at the CDF of standard normal, we get $a = -1.96$, $b = 1.96$. Thus,

$$P\left(-1.96 \leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq 1.96\right) = 0.95$$

With a little rearrangement, we have

$$P\left(\bar{X}_n - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

Therefore, the interval $\left[\bar{X}_n - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X}_n + 1.96\frac{\sigma}{\sqrt{n}}\right]$ contains the true mean 95% of the times.

Theorem 7.8. *The* $100(1-\alpha)\%$ *confidence interval for the sample mean* $\bar{X}_n$ *is* $\bar{X}_n \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$, *where* $z_{\alpha/2}$ *is the critical value such that* $\Phi(z_{\alpha/2}) = \frac{\alpha}{2}$.

In practice, because we do not know $\sigma/\sqrt{n}$, we replace it with the standard error $s/\sqrt{n}$. Thus, we compute the confidence interval as $\bar{X}_n \pm z_{\alpha/2}SE$. However, this replacement is not without risk. When the sample size is small, s is a very poor estimate of σ. For the approximation to be valid, we require either the sample size is large enough ($n \geq 30$ at least) or the population distribution is nearly normal. Some commonly used confidence levels:

- 90% CI: $\alpha = 0.1$, $z_{0.05} = 1.645$
- 95% CI: $\alpha = 0.05$, $z_{0.025} = 1.96$
- 99% CI: $\alpha = 0.01$, $z_{0.005} = 2.58$

We go through some common misunderstandings about confidence intervals through an example. Suppose we have a sample fo size 50 with mean 3.2 and standard deviation 1.74. We construct the 95% confidence interval as

$$\bar{X} \pm 1.96 \times \frac{1.74}{\sqrt{50}} \approx 3.2 \pm 0.5 = (2.7, 3.7).$$

Now check the following interpretations (true or false):

1. We are 95% confident that the sample mean is between 2.7 and 3.7.

   False. The CI definitely contains the sample mean $\bar{X}$.

2. 95% of the population observations are in 2.7 to 3.7.

   False. The CI is about covering the population mean, not for covering 95% of the entire population.

3. The true mean falls in the interval (2.7, 3.7) with probability 95%.

   False. The true mean μ is a fixed number, not a random one that happens with a probability.

4. If a new random sampleis taken, we are 95% confident that the new sample mean will be between 2.7 and 3.7.

   False. The confidence interval is for covering the population mean, not for covering the mean of another sample.

5. This confidence interval is not valid if the population or sample is not normally distributed.

   False. The construction of the CI only uses the normality of the sampling distribution of the sample mean (by the CLT). Neither the population nor the sample is required to be normally distributed.

So what is exactly the thing that has a 95% change to happen? It is the procedure to construct the 95% interval. About 95% of the intervals constructed following the procedure will cover the true population mean μ. After taking the sample and an interval is constructed, the constructed interval either covers μ or it doesn't. But if we were able to take many such samples and reconstruct the interval many times, 95% of the intervals will contain the true mean.

7.6 Hypothesis testing*

Confidence interval allows us to construct an interval estimate of a population parameter. Hypothesis testing allows us to test specific hypothesis about a population parameter with the evidence obtained from a sample. The earliest use of statistical hypothesis testing is generally credited to the question of whether male and female births are equally likely (null hypothesis), which was addressed in the 1700s by John Arbuthnot and later by Pierre-Simon Laplace.

Let p be the population ratio (defined as the ratio of boys to the total number of babies). We hypotheses that

$$H_0 : p = 0.5$$

This is called the **null hypothesis**, which is the hypothesis we want to test. If the null hypothesis is false, we have

$$H_1 : p \neq 0.5$$

This is called the **alternative hypothesis**. How am I able to test which hypothesis is true? I can answer this question by collecting a small sample. Suppose I have collected a sample of 50 babies computed a sample ratio of $\hat{p} = 0.55$. Does it prove or disprove the hypothesis?

Note that the ratio $\hat{p}$ can be regarded as a sample mean. Let X_i be a random variable that equals 1 if the i-th baby is a boy and 0 otherwise. Then, $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$. The variance of $\hat{p}$ is given by

$$Var(\hat{p}) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{p(1-p)}{n}$$

since X_i is a Bernoulli random variable. By the Central Limit Theorem, we have

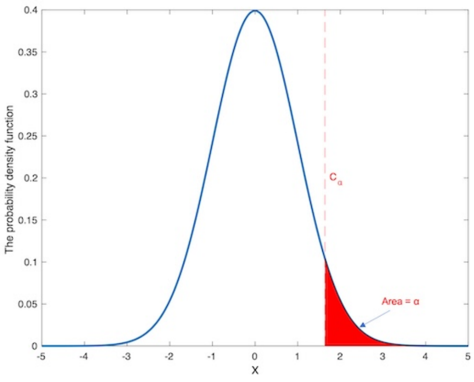
$$\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \to N(0,1)$$

Suppose H_0 is true, then we know the distribution of $\hat{p}$. In particular, there is 95% chance that $\hat{p}$ would be in the interval

$$p \pm 1.96\sqrt{\frac{p(1-p)}{n}} = 0.5 \pm 0.14$$

Our observed sample mean $\hat{p} = 0.55$ is not outrageous. It is well within this interval. That means the evidence is not against the null hypothesis. It does not mean H_0 is true. But it is reasonable given we have observed a sample mean $\hat{p} = 0.55$.

Suppose we have observed $\hat{p} = 0.65$. This piece of evidence does not seem to be consistent with the null hypothesis. Because if H_0 is true, we only have less than 5% chance of observing this sample mean. It is extremely unlikely. Based on this sample, we are more inclined to reject the H_0. Rejecting the null hypothesis does not mean it is false, but it means our evidence does not support this hypothesis.



***p*-value**: the probability of obtaining test results at least as extreme as the result actually observed, under the assumption that the null hypothesis is correct. A very small p-value means that such an extreme observed outcome would be very unlikely under the null hypothesis. Thus, The smaller the p-value, the stronger the evidence against the H_0.

In some studies, we can simply report the p-value and let people judge whether the evidence is strong enough. In other studies, we prefer to select a cut-off value α, call the **significance level**, and follow the rule:

- If the p-value $< \alpha$, reject H_0;
- If the p-value $> \alpha$, do not reject H_0.

Commonly used significance levels: 0.05 and 0.01. And we like to use the word "significant" to describe the test result:

- A test with p-value < 0.05 is said to be (statistically) **significant**;
- A test with p-value < 0.01 is said to be highly **significant**.

When we make a decision about accepting or rejecting a hypothesis, there are chances that we make a mistake. There are two types of mistakes: **Type 1 error** and **Type 2 error**.

		Decision	
		Reject H_0	Fail to reject H_0
Truth	H_0 is true	Type 1 error	✓
	H_0 is false	✓	Type 2 error

Type 1 error is rejecting the H_0 when it is true. **Type 2 error** is failing to reject the H_0 when it is false. Usually, it is more important to control the Type 1 error than the the Type 2 error. That is, we want to minimize the chance of falsely rejecting the null hypothesis.

In the example above, we reject the null hypothesis on the ground that there is only 2.3% of the chance that we could observe this sample. Therefore, the probability of Type 1 error is only 2.3%.

If we make decisions based on a significance level, the significance level is the Type 1 error rate. In other words, when using a 5% significance level, there is 5% chance of making a Type 1 error.

$$P(\text{Type 1 error}|H_0 \text{ is true}) = \alpha$$

This is why we prefer small values of α—smaller α reduces the Type 1 error rate. However, significance level doesn't control Type 2 error rate.

Hypothesis testing with z-statistics

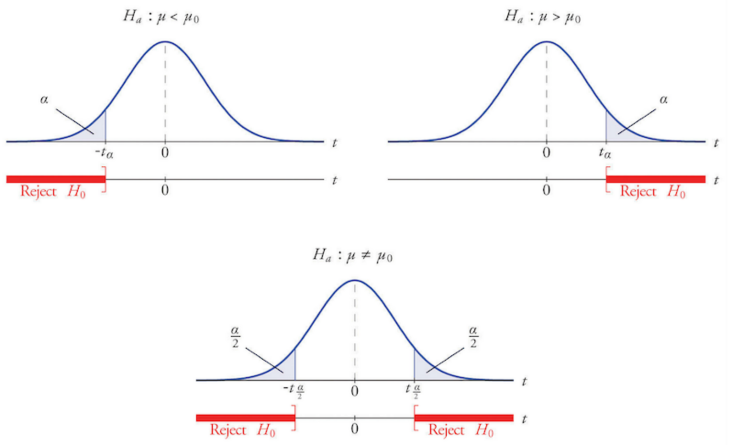
We may have noticed that, in the above example, the assumption that the population σ is known is unrealistic. In practice, we approximate it with the standard error $s/\sqrt{n}$. The approximate is valid if the the sample size is large enough or the underlying distribution is nearly normal. If this is not the case, we would opt for a t-test. Here we summarize the steps of testing for a population mean with z-statistics.

1. Set up the hypothesis:
    - $H_0 : \mu = \mu_0$
    - $H_1 : \mu <$ or $>$ or $\neq \mu_0$
2. Check assumptions and conditions
    - independent and identically distributed ($i.i.d$)
    - Nearly normal distribution or the sample size is large enough
3. Compute the test statistic and the p-value:

$$Z = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

4. Make decision:
    - If p-value $< \alpha$, reject H_0
    - If p-value $> \alpha$, do not reject H_0

We notice that the **two-sided** hypothesis tests are very closed related to the concept of confidence intervals. A two-sided test means we are interested in rejection regions on both sides of the tail distribution. Typically, the alternative hypothesis is $H_1 : \mu \neq \mu_0$.



Suppose we are doing a hypothesis test under the significance level α, the region of accepting the H_0 is

$$-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{SE} \leq z_{\alpha/2}$$

such that the rejection region (p-value) has probability α. This is equivalent to

$$\bar{X} - z_{\alpha/2}SE \leq \mu \leq \bar{X} + z_{\alpha/2}SE$$

which is exactly the $100(1-\alpha)\%$ confidence interval of $\bar{X}$. Therefore, for a two-sided test, we have the rule:

- Reject H_0 if μ is not in the $100(1-\alpha)\%$ CI: $\bar{X} \pm z_{\alpha/2}SE$

We conclude this chapter by reiterating a couple of critical points that could be easily misunderstood.

Rejecting H_0 doesn't means we are 100% sure that H_0 is false. We might make Type 1 errors. Setting a significance level just guarantee we won't make Type 1 error too often.

Failing to reject H_0 does not necessarily mean H_0 is true. We could make a type 2 error when failing to reject H_0. Moreover, unlike type 1 error rate is controlled at a low level, type 2 error rate is usually quite high. When we fail to reject H_0, it just means the data are not able to distinguish between H_0 and

H_1. That's why we say *fail to reject.* p-value is not the probability that the H_0 is true.

Saying that results are statistically significant just informs the reader that the findings are unlikely due to chance alone. However, it says nothing about the practical importance of the finding. For example, rejecting the H_0: $\mu = \mu_0$ does not tell us how big the difference $|\mu - \mu_0|$ is. Mostly in practice we care more about the magnitude of this difference, rather than the fact that they are indeed different. It is possible that the difference is too small to be relevant even if it is significant.

Hypothesis testing with t-statistics

When the sample size is small, we opt for t-test for more reliable hypothsis testing. Define test statistics

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

where s is the sample standard deviation. For small samples, this test statistics follows a Student t-distribution with n degrees of freedom, $T \sim t(n)$.

Why Student-t distribution? Recall the definition of Student-t distribution: when the underlying distribution of $X_1, X_2, \ldots, X_n$ is Normal, sample variance s^2 follows a χ^2 distribution. T follows t distribution by definition regardless of the sample size. However, if the underlying distribution is not normal, this argument loses ground. We use t-test mainly as a convention. But t distribution has heavier tails than standard normal, meaning that we are more likely to reject a hypothesis based on t distribution. In other words, t-test is a more conservative choice than z-test for small samples.

one-tail α	0.05	0.025	0.005
two-tail α	0.10	0.05	0.01
d.f.			
10	1.812	2.228	3.169
20	1.725	2.086	2.845
30	1.697	2.042	2.750
z value	1.645	1.960	2.576

The table shows a few critical values for t-test with different degrees of freedom (d.f.). We can see as the sample size gets larger, t distribution converges to standard normal.