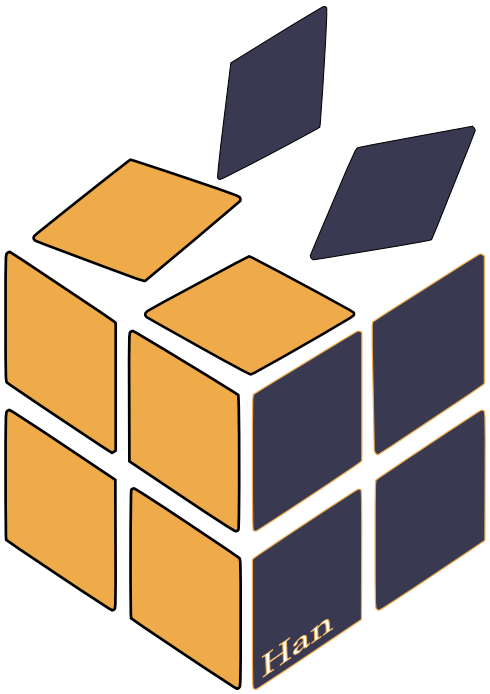
Introduction to Probability and Statistics



Study Note

Han Cheol Moon
School of Computer Science and Engineering
Nanyang Technological University
Singapore
hancheol001@e.ntu.edu.sg
September 27, 2025

Contents

Part I

Introduction to Probability

Chapter 1

Introduction

1.1 Basics

The probability of event A as

$$P(A) = \frac{\text{Number of times A occurs}}{\text{Total number of outcomes}}$$

This commonsense understanding of probability is called the *relative frequency definition.*

Chapter 2

Combinatorics

2.1 Multiplication Principle

Suppose that we perform r experiments such that the k-th experiment has n_k possible outcomes, for $k = 1, 2, \ldots, r$. Then there are a total of $n_1 \times n_2 \times n_3 \times \cdots \times n_r$ possible outcomes for the sequence of r experiments.

2.2 Ordered Sampling with Replacement

Here we have a set with n elements (*e.g.,* $A = \{1, 2, 3, \ldots n\}$), and we want to draw k samples from the set such that ordering matters and repetition is allowed. For example, if $A = \{1, 2, 3\}$ and $k = 2$, there are 9 different possibilities. In general, we can argue that there are$k$ positions in the chosen list: (Position 1, Position 2, ..., Position k). There are n options for each position. Thus, when ordering matters and repetition is allowed, the total number of ways to choose k objects from a set with n elements is

$$n \times n \times \cdots \times n = n^k.$$

2.3 Ordered Sampling without Replacement: Permutations

Consider the same setting as above, but now repetition is not allowed. For example, if $A = \{1, 2, 3\}$ and $k = 2$, there are 6 different possibilities. In general, we can argue that there are k positions in the chosen list: (Position 1, Position 2, ..., Position k). There are n options for the first position, $(n - 1)$ options for the second position (since one element has already been allocated to the first position and cannot be chosen here), $(n - 2)$ options for the third position, and $(n - k + 1)$ options for the k-th position. Thus, when ordering matters and repetition is not allowed, the total number of ways to choose k objects from a set with n elements is

$$n \times (n - 1) \times \cdots \times (n - k + 1).$$

It si called a k permutation of the elements in set A. We use the following notation:

$$P_k^n = n \times (n - 1) \times \cdots \times (n - k + 1).$$

Note that if k is larger than n, then $P_k^n = 0$.

Example: Birthday problem or birthday paradox is a problem that If k people are at a party, what is the probability that at least two of them have the same birthday? Suppose that there are $n = 365$ days in a year and all days are equally likely to be the birthday of a specific person.

$$P(A) = 1 - \frac{P_k^n}{n^k}.$$

The reason this is called a paradox is that $P(A)$ is numerically different from what most people expect. For example, if there are $k = 23$ people in the party, what do you guess is the probability that at least two of them have the same birthday, $P(A)$? The answer is .5073, which is much higher than what most people guess. The probability crosses 99 percent when the number of peoples reaches 57. But why is the probability higher than what we expect?

It is important to note that in the birthday problem, neither of the two people are chosen beforehand. To better answer this question, let us look at a different problem: I am in a party with $k - 1$ people. What is the probability that at least one person in the party has the same birthday as mine? Well, we need to choose the birthdays of $k - 1$ people, the total number of ways to do this is n^{k-1}. The total number of ways to choose the birthdays so that no one has my birthday is $(n - 1)^{k-1}$. Thus, the probability that at least one person has the same birthday as mine is

$$P(B) = 1 - \left(\frac{n-1}{n}\right)^{k-1}.$$

Now, if $k = 23$, this probability is only $P(B) = 0.0586$, which is much smaller than the corresponding $P(A) = 0.5073$. The reason is that event B is looking only at the case where one person in the party has the same birthday as me. This is a much smaller event than event A which looks at all possible pairs of people. Thus, $P(A)$ is much larger than $P(B)$. We might guess that the value of $P(A)$ is much lower than it actually is, because we might confuse it with $P(B)$.

Permutations of n elements: An n-permutation of n elements is just called a permutation of those elements. In this case $k = n$ and we have

$$\begin{aligned} P_n^n &= n \times (n-1) \times \cdots \times (n-n+1) \\ &= n \times (n-1) \times \cdots \times 1, \end{aligned}$$

which is denoted $n!$. We can rewrite as

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

2.4 Unordered Sampling without Replacement: Combinations

Here we have a set with n elements, *e.g.,* $A = \{1, 2, 3, \ldots, n\}$ and we want to draw k samples from the set such that ordering does not matter and repetition is not allowed. Thus, we basically want to choose a k-element subset of A, which we also call a k-combination of the set A. For example if $A = \{1, 2, 3\}$ and $k = 2$, there are 3 different possibilities. We show the number of k-element subsets of A by

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

This is also called the *binomial coefficient*. This is because the coefficients in the binomial theorem are given by

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

An intuitive way to understand this is that there are $n \times (n-1) \times \cdots \times (n-k+1)$ ways to place items and the $k \times \cdots \times 1$ ways to order the times, which can be ignored.

A simple way to find $\binom{n}{k}$ is to compare it with P_k^n. Note that the difference between the two is ordering.

$$P_k^n = \binom{n}{k} \times k!.$$

Example 1: I choose 3 cards from the standard deck of cards. What is the probability that these cards contain at least one ace?

- The sample space contains all possible ways to choose 3 cards from 52 cards.
- There are $52 - 4 = 48$ non-ace cards

Example 2: How many distinct sequences can we make using 3 letter "A"s and 5 letter "B"s? (AAABBBBB, AABABBBB, .)

You can think of this problem in the following way. You have 3+5=8 positions to fill with letters A or B. From these 8 positions, you need to choose 3 of them for "A"s. Whatever is left will be filled with "B"s. Thus the total number of ways is

$$\binom{8}{3}.$$

Equivalently, you chould have chosen the locations for Bs.

$$\binom{8}{5}.$$

The same argument can be repeated for general n and k to conclude

$$\binom{n}{k} = \binom{n}{n-k}.$$

2.5 Bernoulli Trials and Binomial Distribution

A *Bernoulli Trial* is a random experiment that has two possible outcomes which we can label as "success" and "failure", such as

- You toss a coin. The possible outcomes are H and T.

We usually denote the probability of success by p and probability of failure by $q = 1 - p$. If we have an experiment in which we perform n independent Bernoulli trials and count the total number of successes, we call it a binomial experiment. For example, you may toss a coin n times repeatedly and be interested in the total number of heads.

Example: Suppose that I have a coin for which $P(H) = p$ and $P(T) = 1 - p$. I toss the coin 5 times.

- $P(THHHH) = p(T) \times p(H) \cdots = (1-p)p^4$
- $P(HTHHH) = (1-p)p^4$
- $P(HHTHH) = (1-p)p^4$
- $B = \{THHHH, HTHHH, HHTHH, HHHTH, HHHHT\}$, $P(B) = 5p^4(1-p)$
- Let $C = \{TTHHH, THTHH, \dots\}$.

$$\begin{aligned} P(C) &= P(TTHHH) + P(THTHH) + \dots \\ &= |C|p^3(1-p)^2 \end{aligned}$$

- The $|C|$ is the total number of distinct sequences that you can create using two tails and three heads.

$$\binom{5}{3}.$$

- Therefore,

$$P(C) = \binom{5}{3} p^3(1-p)^2$$

Now we can define *Binomial Formula*: For n independent Bernoulli trials where each trial has success probability p, the probability of k successes is given by

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

Similarly, *multinomial coefficients* is given by

$$\binom{n}{n_1, n_2, ..., n_r} = \frac{n!}{n_1! n_2! ... n_r!}.$$

2.6 Unordered Sampling with Replacement

Suppose that we want to sample from the set $A = \{a_1, a_2, \dots, a_n\}$ k times such that repetition is allowed and ordering does not matter. For example, if $A = \{1, 2, 3\}$ and $k = 2$, then there are 6 different ways of doing this.

How can we get the number 6 without actually listing all the possibilities? One way to think about this is to note that any of the pairs in the above list can be represented by the number of 1's, 2's and 3's it contains. That is, if x_1 is the number of ones, x_2 is the number of twos, and x_3 is the number of threes, we can equivalently represent each pair by a vector (x_1, x_2, x_3), *i.e.,*

- $(1,1) \to (x_1, x_2, x_3) = (2, 0, 0)$
- $(1,2) \to (x_1, x_2, x_3) = (1, 1, 0)$
- $(2,3) \to (x_1, x_2, x_3) = (0, 1, 1)$

Note that here $x_i \geq 0$ are integers and $x_1 + x_2 + x_3 = 2$. Thus, we can claim that the number of ways we can sample two elements from the set $A = \{1, 2, 3\}$ such that ordering does not matter and repetition is allowed is the same as solutions to the following equation

$$x_1 + x_2 + x_3 = 2,$$

where $x_i \in \{0, 1, 2\}$. We can generalize this by saying: The total number of distinct k samples from an n-element set such that repetition is allowed and ordering does not matter is the same as the number of distinct solutions to the equation

$$x_1 + x_2 + \cdots + x_n = k,$$

where $x_i \in \{0, 1, 2, \dots\}$. The number of distinct solution to the equation is given by

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

Proof 1 *Let us first define following simple mapping in which we replace an integer x_i with vertical lines* i.e., $|$. *For instance, $x_1 + x_2 + x_3 = 2$, then we can equivalently write $|++|$ for $1 + 0 + 1$. We have an unique representation using vertical lines and plus signs. Each solution can be represented by k vertical lines and $n - 1$ plus sings. Thus, we get*

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

Chapter 3

Discrete Random Variables

3.1 Random Variables

A random variable X is a function from the sample space to the real numbers.

$$X : S \to \mathbb{R}$$

3.2 Probability Mass Function (PMF)

If X is a discrete random variable then its range R_X is a countable set, so, we can list the elements in R_X. In other words, we can write

$$R_X = \{x_1, x_2, \dots\}$$

Note that here $x_1, x_2, \dots$ are possible values of the random variable X. While random variables are usually denoted by capital letters, to represent the numbers in the range we usually use lowercase letters. For a discrete random variable X, we are interested in knowing the probabilities of $X = x_k$.

Let X be a discrete random variable with range $R_X = \{x_1, x_2, \dots\}$ (finite or countably infinite). The function

$$P_X(x_k) = P(X = x_k), \text{ for } k = 1, 2, 3 \dots$$

is called the *probability mass function* (PMF) of X. Why is it called PMF? In physics, mass is the weight over gravity:

$$m = \frac{W}{g}$$

In statistics, the probability of a discrete random variable is:

$$P(A) = \frac{n(A)}{n(all)}.$$

Thus, the weight (W) is analogous to the number of ways an event A can occur ($n(A)$) and the gravity is analogous to the sample space ($n(all)$).

3.3 Special Distributions

3.3.1 Bernoulli Distribution

$$P_X(x) = \begin{cases} p & \text{for } x = 1 \\ 1-p & \text{for } x = 0 \\ 0 & \text{Otherwise} \end{cases}$$

A Bernoulli random variable is associated with a certain event A. If event A occurs (for example, if you pass the test), then $X = 1$; otherwise $X = 0$. For this reason the Bernoulli random variable, is also called the *indicator* random variable.

3.3.2 Geometric Distribution

Suppose that I have a coin with $P(H) = p$. I toss the coin until I observe the first heads. We define X as the total number of coin tosses in this experiment. Then X is said to have geometric distribution with parameter p. In other words, you can think of this experiment as **repeating independent Bernoulli trials until observing the first success**. The range of X here is $RX = \{1, 2, 3, \dots\}$.

$$P_X(k) = P(X = k) = (1-p)^{k-1}p, \text{ for } k = 1, 2, 3, \dots$$

3.3.3 Binomial Distribution

Suppose that I have a coin with $P(H) = p$. I toss the coin n times and define X to be the total number of heads that I observe. Then X is binomial with parameter n and p. The range of X in this case is $R_X = 0, 1, 2, \dots, n$.

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

Here is a useful way of thinking about a binomial random variable. It can be obtained by n independent coin tosses. If we think of each coin toss as a Bernoulli random variable, the $Binomial(n, p)$ random variable is a sum of n independent $Bernoulli(p)$ random variables. This is stated more precisely in the following lemma.

If $X_1, X_2, \dots, X_n$ are independent $Bernoulli(p)$ random variables, then the random variable X defined by $X = X_1 + X_2 + \cdots + X_n$ has a $Binomial(n, p)$ distribution.

Example:

- Let $X \sim Binomial(n, p)$ and $Y \sim Binomial(m, p)$ be two independent random variables. Define a new random variable as $Z = X + Y$. Find the PMF of Z.
- Solution 1: Since $X \sim Binomial(n, p)$, we can think of X as the number of heads in n independent coin tosses:

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

where the X_i's are independent Bernoulli RVs. Similarly, $Y \sim Binomial(m, p)$. Thus, the RV $Z = X + Y$ will be the total number of heads in $n + m$ coin tosses:

$$Z = X + Y = X_1 + \cdots + X_n + Y_1 + \cdots + Y_m.$$

Therefore, Z is a binomial RV with parameters $m + n$ and p, *i.e.,* $Binomial(m + n, p)$.

- Solution 2: First, we note that $R_z = \{0, 1, \ldots, m + n\}$. For $k \in R_z$, we get

$$P_Z(k) = P(Z = k) = P(X + Y = k).$$

We will find $P(X + Y = k)$ by using conditioning and the law of total probability.

$$\begin{aligned}
P(Z = k) &= P(X + Y = k) \\
&= \sum_{i=0}^{n} P(X + Y = k | X = i) P(X = i) \\
&= \sum_{i=0}^{n} P(Y = k - i | X = i) P(X = i) \\
&= \sum_{i=0}^{n} P(Y = k - i) P(X = i) \quad \text{Since } X \text{ and } Y \text{ are independent} \\
&= \sum_{i=0}^{n} \binom{m}{k-i} p^{k-i} (1-p)^{m-k+i} \binom{n}{i} p^{i} (1-p)^{n-i} \\
&= \sum_{i=0}^{n} \binom{m}{k-i} \binom{n}{i} p^{k} (1-p)^{m+n-k} \\
&= p^{k} (1-p)^{m+n-k} \sum_{i=0}^{n} \binom{m}{k-i} \binom{n}{i} \\
&= \binom{m+n}{k} p^{k} (1-p)^{m+n-k} \quad \text{by Vandermonde's identity}
\end{aligned}$$

3.3.4 Binomial v.s., Multinomial Distributions

Binomial Distribution models **the number of successes in a fixed number of independent trials, where each trial has only two possible outcomes**: success (with probability p) or failure (with probability $1 - p$).

- Number of outcomes per trial: Two (success/failure).
- Trials: Fixed and independent.
- Parameters: n (number of trials) and p (probability of success in a single trial).
- Probability Mass Function (PMF):

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

where k is the number of successes, $0 \leq k \leq n$.

- Example: Flipping a coin $n = 10$ times and counting the number of heads (successes).

Multinomial Distribution generalizes the binomial distribution to experiments with more than two possible outcomes per trial. It models the counts of each outcome across a fixed number of trials.

- Number of outcomes per trial: More than two (e.g., k categories).
- Trials: Fixed and independent.
- Parameters: n (number of trials) and $\mathbf{p} = (p_1, p_2, \ldots, p_k)$, where p_i is the probability of outcome i and $\sum_{i=1}^{k} p_i = 1$.
- Probability Mass Function (PMF):

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = \frac{n!}{x_1! x_2! \cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k},$$

  where x_i is the count of outcomes in category i, and $\sum_{i=1}^{k} x_i = n$.
- Example: Rolling a six-sided die $n = 10$ times and counting how many times each face $(1, 2, \ldots, 6)$ appears.

Negative Binomial (Pascal) Distribution The *negative binomial* or *Pascal distribution* is a **generalization of the geometric distribution**. It relates to the random experiment of **repeated independent trials until observing m successes**. Suppose that I have a coin with $P(H) = p$. I toss the coin until I observe m heads, where $m \in \mathbb{N}$. We define X as the total number of coin tosses in this experiment. Then X is said to have Pascal distribution with parameter m and p. We write $X \sim Pascal(m, p)$. Note that $Pascal(1, p = Geometric(p))$, since the geometric distribution repeats trials until observing the first success. Note that by our definition the range of X is given by $R_X = \{m, m+1, m+2, m+3, \ldots\}$, since X is the number of coin tosses to observe m target events.

Let's derive the PMF of a $Pascal(m, p)$ RV X. To find the probability of the event $A = \{X = k\}$, we argue as follows. By definition, event A can be written as $A = B \cap C$, where

- B is the event that we observe $m - 1$ heads (*i.e.,* successes) in the first $k - 1$ trials
- C is the event that we observe a head in the k-th trial.

Note that B and C are independent events because they are related to different independent trials (coin tosses). Thus,

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

We get $P(C) = p$, so

$$P(B) = \binom{k-1}{m-1} p^{m-1} (1-p)^{(k-1)-(m-1)} = \binom{k-1}{m-1} p^{m-1} (1-p)^{k-m}.$$

Finally, we obtain

$$P(B) = \binom{k-1}{m-1} p^{m} (1-p)^{k-m}.$$

3.3.5 Hyper-Geometric Distribution

You have a bag that contains b blue marbles and r red marbles. You choose $k \leq b+r$ marbles at random (without replacement). Let X be the number of blue marbles in your sample. By this definition, we have $X \leq \min(k, b)$. Also, the number of red marbles in your sample must be less than or equal to r, so we conclude $X \geq \max(0, k-r)$. Therefore, the range of X is given by $R_X = \{\max(0, k-r), \max(0, k-r)+1, \max(0, k-r)+2, \ldots, \min(k, b)\}$.

To find $P_X(x)$, note that total number of ways to choose k marbles from $b+r$ marbles is $\binom{b+r}{k}$. The total number of ways to choose x blue marbles and $k-x$ red marbles is $\binom{b}{x}\binom{r}{k-x}$. Thus, we get

$$P_X(x) = \frac{\binom{b}{x}\binom{r}{k-x}}{\binom{b+r}{k}}, \quad \text{for } x \in R_X.$$

3.3.6 Poisson Distribution

The Poisson distribution is one of the most widely used probability distributions. It is usually used in scenarios where we are **counting the occurrences of certain events in an interval of time or space**. In practice, it is often an approximation of a real-life random variable. Here is an example of a scenario where a Poisson random variable might be used. Suppose that we are counting the number of customers who visit a certain store from 1pm to 2pm. Based on data from previous days, we know that on average $\lambda = 15$ customers visit the store. Of course, there will be more customers some days and fewer on others. Here, we may model the random variable X showing the number customers as a Poisson random variable with parameter $\lambda = 15$. Let us introduce the Poisson PMF first, and then we will talk about more examples and interpretations of this distribution.

$$P_X(k) = e^{-\lambda}\frac{\lambda^k}{k!}.$$

Note that λ is the mean number of events within a given interval of time or space.

Example: The number of emails that I get in a weekday can be modeled by a Poisson distribution with an average of 0.2 emails per minute.

- What is the probability that I get no emails in an interval of length 5 minutes?
  - For 5 minutes, there would be 1 email on average. Thus, $\lambda = 1$,

$$P(X=0) = P_X(0) = e^{-\lambda}\frac{\lambda^0}{0!} = \frac{1}{e} \approx 0.37$$

- What is the probability that I get more than 3 emails in an interval of length 10 minutes?
  - Let Y be the number of emails that I get in the 10-minute interval. Then by the assumption Y is a Poisson RV with parameter $\lambda = 10 \times 0.2 = 2$. Thus,

$$\begin{aligned} P(Y>3) &= 1 - P(Y \leq 3) \\ &= 1 - (P_Y(0) + P_Y(1) + P_Y(2) + P_Y(3)) \\ &= 1 - e^{-\lambda} - \frac{e^{-\lambda}\lambda}{1!} - \frac{e^{-\lambda}\lambda^2}{2!} - \frac{e^{-\lambda}\lambda^3}{3!} \\ &\approx 0.1429 \end{aligned}$$

Imagine you have a busy customer service center that receives phone calls. You want to know how many calls to expect in an hour, but calls can come at any moment and don't follow a strict schedule.

- Average Rate (λ): First, you determine the average number of calls you receive per hour. Let's say it's 10 calls per hour. This average rate is denoted by the symbol λ
- Probability Calculation: Using the Poisson formula, you can calculate the probability of receiving a certain number of calls in any given hour.

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

- $P(X = k)$ is the probability of getting k calls in an hour.
- e is the base of the natural logarithm (approximately equal to 2.71828).
- λ is the average rate (10 calls per hour).
- k is the number of calls you want to find the probability for.
- $k!$ (k factorial) is the product of all positive integers up to k.

3.3.7 Poisson as an Approximation for Binomial

The Poisson distribution can be viewed as the limit of binomial distribution. Suppose $X \sim Binomial(n, p)$ where the number of trials n is very large and the probability of success p is very small. In particular, assume that $\lambda = np$ is a positive constant. We show that the PMF of X can be approximated by the PMF of a $Poisson(\lambda)$ random variable. The importance of this is that Poisson PMF is much easier to compute than the binomial. Let us state this as a theorem.

Let $X \sim Binomial(n, p = \frac{\lambda}{n})$, where $\lambda > 0$ is fixed. Then for any $k \in \{0, 1, 2, \dots\}$ we have

$$\lim_{n\to\infty} P_X(k) = \frac{e^{-\lambda}\lambda^k}{k!}.$$

References Poisson

3.4 Cumulative Distribution Function

The PMF is one way to describe the distribution of a discrete RV. As we will see later on, PMF cannot be defined for continuous random variables. The cumulative distribution function (CDF) of a random variable is another method to describe the distribution of random variables. The advantage of the CDF is that it can be defined for any kind of RV (discrete, continuous, and mixed).

Definition 1 *Cumulative Distribution Function* The cumulative distribution function (CDF) of random variable X is defined as

$$F_X(x) = P(X \leq x), \forall x \in \mathbb{R}.$$

Note that the subscript X indicates that this is the CDF of the random variable X. Also, note that the CDF is defined for all $x \in \mathbb{R}$.

3.5 Expectation

If you have a collection of numbers $a_1, a_2, \ldots, a_N$, their average is a single number that describes the whole collection. Now, consider a random variable X. We would like to define its average, or as it is called in probability, its expected value or mean. The expected value is defined as the weighted average of the values in the range.

Definition 2 *Expected Value* Let X be a discrete RV with range $R_X = \{x_1, x_2, \ldots, \}$. The expected value of X, denoted by EX is defined as

$$EX = \sum_{x_k \in R_X} x_k P(X = x_k) = \sum_{x \in R_X} x_k P_X(x_k)$$

3.6 Functions of Random Variables

If X is a RV and $Y = g(X)$, then Y itself is a random variable. Thus, we can talk about its PMF, CDF, and expected value. First note that the range of Y can be written as

$$R_Y = \{g(x) | x \in R_X\}.$$

If we already know the PMF of X, to find the PMF of $Y = g(X)$, we can write

$$\begin{aligned} P_Y(y) &= P(Y = y) \\ &= P(g(X) = y) \\ &= \sum_{x: g(x) = y} P_X(x) \end{aligned}$$

Example: Let X be a discrete RV with $P_X(k) = \frac{1}{5}$ for $k = -1, 0, 1, 2, 3$. Let $Y = 2|X|$. Find the range and PMF of Y. The range of Y is

$$\begin{aligned} R_Y &= \{2|x|\} \\ &= \{0, 2, 4, 6\} \end{aligned}$$

To find $P_Y(y)$, we need to find $P(Y = y)$ for $y = 0, 2, 4, 6$:

$$\begin{aligned} P_Y(0) &= P(Y = 0) = P(2|x| = 0) \\ &= P(X = 0) = \frac{1}{5} \\ P_Y(2) &= P(Y = 2) = P(2|x| = 2) \\ &= P(X = -1 \text{ or } X = 1) \\ &= P_X(-1) + P_X(1) = \frac{2}{5} \\ &\vdots \end{aligned}$$

3.6.1 Expected Value of a Function of a Random Variable (LOTUS)

Let X be a discrete random variable with PMF $P_X(x)$, and let $Y = g(X)$. Suppose that we are interested in finding EY. One way to find EY is to first find the PMF of Y and then use the expectation formula $EY = E[g(X)] = \sum_{y \in R_y} y P_Y(y)$. But there is another way which is usually easier. It is called the *law of the unconscious statistician* (LOTUS).

$$\mathbb{E}[g(X)] = \sum_{x_k \in R_X} g(x_k) P_X(x_k)$$

One of the main points of the theorem is that you can compute $\mathbb{E}[g(X)]$ without computing $P_Y(y)$. In practice it is usually easier to use LOTUS than direct definition when we need $\mathbb{E}[g(X)]$.

3.7 Variance

The variance of a random variable X, with mean $EX = \mu_X$, is defined as

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

To compute $\text{Var}(X) = \mathbb{E}[(X - \mu_X)^2]$, note that we need to find the expected value of $g(X) = (X - \mu_X)^2$, so we can use **LOTUS**. In particular, we can write

$$\text{Var}(X) = \mathbb{E}[(X - \mu_X)^2] = \sum_{x_k \in R_X} (x_k - \mu_X)^2 P_X(x_k).$$

3.8 Standard Deviation

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

A useful formula for computing the variance is

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

We can find $\mathbb{E}[X^2]$ using LOTUS:

$$\mathbb{E}[X^2] = \sum_{x_k \in R_X} x_k^2 P_X(x_k).$$

Chapter 4

Continuous and Mixed Random Variables

4.1 Introduction

Remember that discrete random variables can take only a countable number of possible values. On the other hand, a continuous random variable X has a range in the form of an interval or a union of non-overlapping intervals on the real line (possibly the whole real line). Also, for any $x \in \mathbb{R}$, $P(X = x) = 0$. Thus, we need to develop new tools to deal with continuous random variables. The good news is that the theory of continuous random variables is completely analogous to the theory of discrete random variables. Indeed, if we want to oversimplify things, we might say the following: take any formula about discrete random variables, and then replace sums with integrals, and replace PMFs with probability density functions (PDFs), and you will get the corresponding formula for continuous random variables.

4.2 Probability Density Function (PDF)

To determine the distribution of a discrete random variable we can either provide its PMF or CDF. For continuous random variables, the CDF is well-defined so we can provide the CDF. However, **the PMF does not work for continuous random variables, because for a continuous random variable** $P(X = x) = 0$, $\forall x \in \mathbb{R}$. Instead, we can usually define the *probability density function* (PDF). The PDF is the *density* of probability rather than the probability mass. The concept is very similar to mass density in physics: its unit is probability per unit length. For example, let the bus waiting time be uniformly distributed: $X \sim [10, 30]$. The probability of waiting between 15 and 20 minutes is:

$$P(X < 15) = \int_{15}^{20} \frac{1}{20}\, dx = \frac{1}{20} \cdot (20 - 15) = \frac{1}{4} = 0.25.$$

So, the mass is analogous to the interval ($[a, b] = [15, 20]$) and the volume is analogous to the entire range ($[c, d] = [10, 30]$). To get a feeling for PDF, consider a continuous random variable X and define the function $f_X(x)$ as follows (wherever the limit exists):

$$f_X(x) = \lim_{\Delta \to 0^+} \frac{P(x < X \leq x + \Delta)}{\Delta}.$$

The function $f_X(x)$ gives us the probability density at point x. It is the limit of the probability of the interval $(x, x+\Delta]$ divided by the length of the interval as the length of the interval goes to 0. Remember that

$$P(x < X \leq x + \Delta) = F_X(x + \Delta) - F_X(x).$$

Thus, we get

$$\begin{aligned} f_X(x) &= \lim_{\Delta \to 0} \frac{F_X(x+\Delta) - F_X(x)}{\Delta} \\ &= \frac{dF_X(x)}{dx} \\ &= F'_X(x), \quad \text{if } F_X(x) \text{ is differentiable at } x. \end{aligned}$$

Let's find the PDF of the uniform random variable $X \sim Uniform(a, b)$, which can be expressed as follows:

$$f_X(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & x < a \text{ or } x > b \end{cases}$$

Note that the CDF is not differentiable at points a and b. Nevertheless, this is not important at this moment.

The uniform distribution is the simplest continuous random variable you can imagine. For other types of continuous random variables the PDF is non-uniform. Note that for small values of δ we can write

$$P(x < X \leq x + \delta) \approx f_X(x)\delta.$$

Thus, if $f_X(x_1) < f_X(x_2)$, we can say $P(x_1 < X \leq x_1 + \delta) < P(x_2 < X \leq x_2 + \delta)$, *i.e.,* the value of X is more likely to be around x_2 than x_1.

Since the PDF is the derivative of the CDF, the CDF can be obtained from PDF by integrations (by assuming absolute continuity):

$$F_X(x) = \int_{-\infty}^{x} f_X(u)du.$$

Also, we have

$$P(a < X \leq b) = F_X(b) - F_X(a) = \int_a^b f_X(u)du.$$

More generally, for a set A, $P(X \in A) = \int_a^b f_X(u)du$. Note that if we integrate over the entire real line, we must get 1, *i.e.,*

$$\int_{-\infty}^{\infty} f_X(u)du = 1.$$

4.3 Expected Value and Variance

As we mentioned earlier, the theory of continuous random variables is very similar to the theory of discrete random variables. In particular, usually summations are replaced by integrals and PMFs are replaced by PDFs. The proofs and ideas are very analogous to the discrete case, so sometimes we state the results without mathematical derivations for the purpose of brevity.

Recall that the expected value of a discrete random variable can be obtained as

$$EX = \sum_{x_k \in R_X} x_k P_X(x_k).$$

The expected value of a continuous RV as

$$EX = \int_{-\infty}^{\infty} x f_X(x) dx.$$

4.3.1 Expected Value of a Function of a Continuous Random Variable

Law of the unconscious statistician (LOTUS) for continuous random variables:

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

4.3.2 Variance

$$\begin{aligned} \text{Var}(X) &= \mathbb{E}[(X - \mu_X)^2] \\ &= \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x) dx \\ &= EX^2 - (EX)^2 \\ &= \int_{-\infty}^{\infty} x^2 f_X(x) dx - \mu_X^2 \end{aligned}$$

Note that for $a, b \in \mathbb{R}$, we always have

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

4.4 Functions of Continuous Random Variables

If X is a continuous random variable and $Y = g(X)$ is a function of X, then Y itself is a random variable. Thus, we should be able to find the CDF and PDF of Y. It is usually more straightforward to start from the CDF and then to find the PDF by taking the derivative of the CDF. Note that before differentiating the CDF, we should check that the CDF is continuous. As we will see later, the function of a continuous random variable might be a non-continuous random variable. Let's look at an example.

Example: Let X be a $Uniform(0, 1)$ random variable, and let $Y = e^X$.

- CDF of Y
- PDF of Y
- EY

The PDF of X is given by

$$F_X(x) = \begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } 0 \leq x \leq 1 \\ 1 & \text{for } x > 1 \end{cases}$$

The range of x, $R_X = [0, 1]$, so the range of Y, $R_Y = [1, e]$. We can find the CDF of Y as follows:

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P(e^X \leq y) \\ &= P(X \leq \ln y) \quad \text{, since } e^x \text{ is an increasing function.} \\ &= F_X(\ln y) \quad \text{by definition.} \\ &= \ln y \quad \text{, since } F_X(x) = x \text{ for } 0 \leq x \leq 1 \text{ and } 0 \leq \ln y \leq 1. \end{aligned}$$

4.5 The Method of Transformations of Random Variables

So far, we have discussed how we can find the distribution of a function of a continuous random variable starting from finding the CDF. If we are interested in finding the PDF of $Y = g(X)$, where $g(\cdot)$ is some deterministic transformation of X, and the function g satisfies following properties, we can utilize a method called the method of transformations.

- $g(x)$ is differentiable;
- $g(x)$ is a strictly (or monotonically) increasing function, that is, if $x_1 < x_2$, then $g(x_1) < g(x_2)$.

Now, let X be a continuous random variable and $Y = g(X)$. We will show that you can directly find the PDF of Y using the following formula.

$$f_Y(y) = \begin{cases} \frac{f_X(x_1)}{g'(x_1)} = f_X(x_1).\frac{dx_1}{dy} & \text{where } g(x_1) = y \\ 0 & \text{if } g(x) = y \text{ does not have a solution} \end{cases}$$

Note that the derivative $\frac{dx}{dy}$ or $\frac{d}{dy}(g^{-1}(y))$ **measures how X changes with respect to Y.** Since g is strictly increasing, its inverse function g^{-1} is well defined. You can imagine a simple function like a linear function, *e.g.,* $Y = 3X + 1$. Then, for each $y \in R_Y$, there exists a **unique** x_1 such that $g(x_1) = y$. We can write $x_1 = g^{-1}(y)$.

$$\{Y \leq y\} = \{g(X) \leq y\} = \{X \leq g^{-1}(x)\}.$$

Thus,

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P(g(X) \leq y) \\ &= P(X \leq g^{-1}(y)) \quad \text{, since } g \text{ is strictly increasing.} \\ &= F_X(g^{-1}(y)). \end{aligned}$$

To find the PDF of Y, we differentiate $F_Y(y)$ as follows:

$$\begin{aligned} f_Y(y) &= \frac{d}{dy} F_X(x_1) \quad \text{by } g(x_1) = y \\ &= \frac{dx_1}{dy} \cdot \underbrace{\frac{d}{dx_1} F_X(x_1)}_{=F'_X(x_1)} \\ &= \frac{dx_1}{dy} f_X(x_1) \\ &= f_X(g^{-1}(y)) \left| \frac{d}{dy}(g^{-1}(y)) \right| \end{aligned}$$

We can repeat the same argument for the case where g is **strictly decreasing**. In that case, $g'(x_1)$ will be **negative**, so we need to use $|g'(x_1)|$. Thus, we can state the following theorem for a *strictly monotonic function.* (A function $g : R \to R$ is called strictly monotonic if it is strictly increasing or strictly decreasing.)

Actually, we assumed that g was one-to-one out of convenience: the condition that g is one-to-one is not necessary for change of variables to work: Consider a continuous random variable X with domain R_X, and let $Y = g(X)$. Suppose that we can partition R_X into a finite number of intervals such that $g(x)$ is strictly monotone and differentiable on each partition. Then the PDF of Y is given by

$$f_Y(y) = \sum_{i=1}^{n} \frac{f_X(x_i)}{|g'(x_i)|} = \sum_{i=1}^{n} f_X(x_i). \left| \frac{dx_i}{dy} \right|,$$

where $x_1, \ldots, x_n$ are real solutions to $g(x) = y$.

4.5.1 Intuitive Explanation

How to derive the PDF of the random variable $Y = g(X)$ when one knows the PDF of the random variable X? If X is discrete, we can derive the pmf for Y by simply summing up the probability mass for all the x's such that $f(x) = y$. For a general function g, there is no direct formula to get the PDF of the random variable $Y = g(X)$ knowing $p(X)$. There is a formula in case when h is a differentiable one-to-one mapping from the range (*i.e.,* the support) of X to the range of Y.

Take for example a random variable $X \sim \mathcal{N}(\mu, \sigma)$ and set $Y = \exp(X)$. The figure below shows some simulations of X and the corresponding values of Y. The density of X is shown in blue and the one of Y is shown in orange in the vertical direction. Now the question is: knowing the density of X, what is the density of Y? Taking a point y in the range of Y, the PDF f_Y provides the probability of Y, belong to a small area dy around y by the formula below

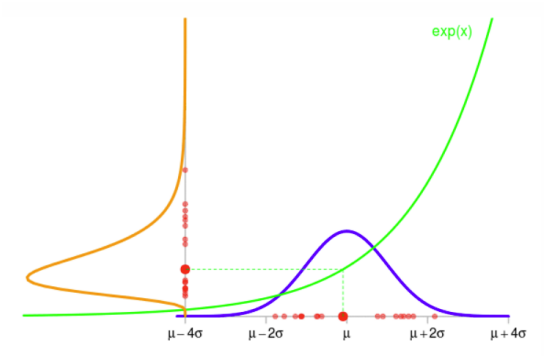
$$P(Y \in dy) \approx f_Y(y)|dy|,$$

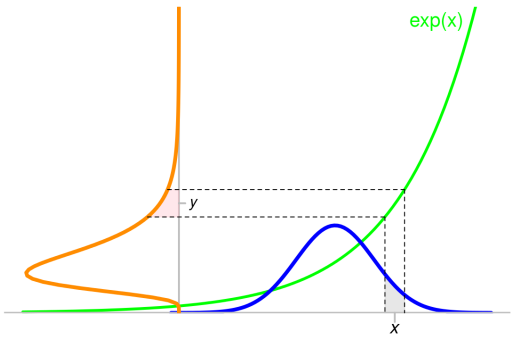
where $P(Y \in dy)$ is the area below the curve. Similarly, we can define

$$P(X \in dx) \approx f_X(x)|dx|$$

The above two areas are approximately the same in case of very small region. Note that if dy and dx are very small, we can approximate the derivative of $g'(x) = \frac{|dy|}{|dx|}$. Compactly, this can be expressed as follows:

$$P(X \in dx) = f_X(x) \frac{|dy|}{g'(x)}$$





With $y = g(x)$ we can get

$$\begin{aligned} P(X \in dx) \approx P(Y \in dy) &= f_X(x)\frac{|dy|}{g'(x)} \\ &= f_X(g^{-1}(y))\frac{|dy|}{g'(g^{-1}(y))} \\ &= f_X(g^{-1}(y))|dy|(g^{-1})'(y) \end{aligned}$$

The last line is by the derivative of inverse function which is

$$\frac{d}{dx}f^{-1}(x) = \frac{1}{f'(f^{-1}(x))}$$

Finally, we can get

$$f_Y(y) = f_X(g^{-1}(y))|(g^{-1})'(y)|$$

Note that the absolute is determined by the function h. This is the so-called *change of variables formula*.

4.6 Various Distributions

4.6.1 Uniform Distribution

$$f_X(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & x < a \text{ or } x > b \end{cases}$$

$$P(c \le x \le d) = \int_c^d f(x)dx = \int_c^d \frac{1}{b-a}dx = \frac{d-c}{b-a}$$

The expected value of a uniform distribution is

$$EX = \int_c^d xf(x)dx = \int_c^d \frac{x}{b-a}dx = \frac{b-a}{2}$$

The variance of a uniform distribution is given by

$$\begin{aligned} \text{Var}(X) &= EX^2 - E^2X \\ &= \int_c^d \frac{x^2}{b-a} - \left(\frac{b-a}{2}\right)^2 dx \\ &= \frac{(b-a)^2}{12} \end{aligned}$$

4.6.2 Exponential Distribution

Check more in Exponential Dist.

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

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

4.6.3 Gamma Distribution

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

Gamma function $\Gamma(x)$ is an extension of the factorial function to real (and complex) numbers. In specific, if $n \in \{1, 2, 3, \dots\}$, then

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

More generally, for any positive real number α, $\Gamma(\alpha)$ is defined as follows:

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

Note that for $\alpha = 1$,

$$\Gamma(\alpha) = 1.$$

Gamma Distribution is a distribution with parameters $\alpha > 0$ and $\lambda > 0$. its PDF is given by

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

4.7 Mixed Random Variables

The mixed random variables are random variables that are **neither discrete nor continuous, but are a mixture of both**.

To find the cumulative distribution function (CDF) of Y, given that $Y = g(X)$ and the transformation $g(X)$ is defined as:

$$g(X) = \begin{cases} X^2 & 0 \le X \le \frac{1}{2} \\ 2X - 1 & \frac{1}{2} < X \le 1 \end{cases}$$

1. Determine the PDF of X. The given PDF of X is:

$$f_X(x) = \begin{cases} 2x & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

2. Determine the ranges of Y. The ranges of Y are derived from the transformation:

   - For $0 \leq X \leq \frac{1}{2}$:

$$Y = X^2$$
$$0 \leq Y \leq \left(\frac{1}{2}\right)^2 = \frac{1}{4}$$

   - For $\frac{1}{2} < X \leq 1$:

$$Y = 2X - 1$$
$$2\left(\frac{1}{2}\right) - 1 < Y \leq 2(1) - 1$$
$$0 < Y \leq 1$$

3. Combining these, we get the range of Y as $0 \leq Y \leq 1$. Find the CDF of Y. The CDF of Y, $F_Y(y)$, is given by $F_Y(y) = P(Y \leq y)$. We need to consider the two different transformations:

   (a) For $0 \leq y \leq \frac{1}{4}$:

$$Y = X^2$$
$$P(Y \leq y) = P(X^2 \leq y) = P(X \leq \sqrt{y})$$
$$F_Y(y) = P(X \leq \sqrt{y}) = \int_0^{\sqrt{y}} 2x\, dx$$
$$F_Y(y) = \left[x^2\right]_0^{\sqrt{y}} = (\sqrt{y})^2 = y$$

   (b) For $\frac{1}{4} < y \leq 1$:

$$Y = 2X - 1$$
$$P(Y \leq y) = P(2X - 1 \leq y) = P(X \leq \frac{y+1}{2})$$
$$F_Y(y) = P(X \leq \frac{y+1}{2}) = \int_0^{\frac{y+1}{2}} 2x\, dx$$
$$F_Y(y) = \left[x^2\right]_0^{\frac{y+1}{2}} = \left(\frac{y+1}{2}\right)^2$$
$$F_Y(y) = \frac{(y+1)^2}{4}$$

4. Combining these results, the CDF of Y is:

$$F_Y(y) = \begin{cases} y & 0 \leq y \leq \frac{1}{4} \\ \frac{(y+1)^2}{4} & \frac{1}{4} < y \leq 1 \end{cases}$$

4.7.1 Delta Function

In this section, we will **use the Dirac delta function to analyze mixed random variables**. Technically speaking, *the Dirac delta function is not actually a function.* It is what we may call a generalized function. Nevertheless, its definition is intuitive and it simplifies dealing with probability distributions.

Remember that any random variable has a CDF. Thus, we can use the CDF to answer questions regarding discrete, continuous, and mixed random variables. On the other hand, the PDF is defined only for continuous random variables, while the PMF is defined only for discrete random variables. Using **delta functions will allow us to define the PDF for discrete and mixed random variables.** Thus, it allows us to unify the theory of discrete, continuous, and mixed random variables.

Dirac Delta Function We cannot define the PDF for a discrete random variable because its CDF has jumps. If we could somehow differentiate the CDF at jump points, we would be able to define the PDF for discrete random variables as well. This is the idea behind our effort in this section. Here, we will introduce the *Dirac delta function* and discuss its application to probability distributions. Let's derive the Dirac delta function.

First, consider the following unit step function $u(x)$:

$$u(x) = \begin{cases} 1 & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

This function has a discontinuity at $x = 0$. Let us remove the jump and define, for any $\alpha > 0$, the function $u_\alpha(x)$ as

$$u_\alpha(x) = \begin{cases} 1 & x > \frac{\alpha}{2} \\ \frac{1}{\alpha}(x + \frac{\alpha}{2}) & -\frac{\alpha}{2} \leq x \leq \frac{\alpha}{2} \\ 0 & x < -\frac{\alpha}{2} \end{cases}$$

The good thing about $u_\alpha(x)$ is that it is a continuous function. Now let us define the function $\delta_\alpha(x)$ as the derivative of $u_\alpha(x)$ wherever it exists.

$$\delta_\alpha(x) = \frac{du_\alpha(x)}{dx} = \begin{cases} \frac{1}{\alpha} & |x| < \frac{\alpha}{2} \\ 0 & |x| > \frac{\alpha}{2} \end{cases}$$

We can notice that

$$\delta_\alpha(x) = \frac{d}{dx}u_\alpha(x), \quad u(x) \overset{\text{a.e.}}{=} \lim_{\alpha \to 0} u_\alpha(x)^{1}$$

Now, we would like to define the delta "function", $\delta(x)$, as

$$\delta(x) = \lim_{\alpha \to 0} \delta_\alpha(x).$$

[1]The term almost everywhere is abbreviated a.e.; in older literature p.p. is used, to stand for the equivalent French language phrase presque partout.

Note that as α becomes smaller and smaller, the height of $\delta_\alpha(x)$ becomes larger and larger and its width becomes smaller and smaller. Taking the limit, we obtain

$$\delta(x) = \begin{cases} \infty & x = 0 \\ 0 & \text{otherwise} \end{cases}$$

Equivalently,

$$\delta(x) = \frac{d}{dx}u(x).$$

Intuitively, with extremely small α, we would like to have the following definitions. Let $g : \mathbb{R} \to \mathbb{R}$ be a continuous function. We define

$$\int_{-\infty}^{\infty} g(x)\delta(x - x_0)dx = \lim_{\alpha \to 0}\left[\int_{-\infty}^{\infty} g(x)\delta_\alpha(x - x_0)dx\right]$$

Then, we have the following lemma, which in fact is the most useful property of the delta function.

Let $g : \mathbb{R} \to \mathbb{R}$ be a continuous function. We have

$$\int_{-\infty}^{\infty} g(x)\delta(x - x_0)dx = g(x_0).$$

Using the Delta Function in PDFs of Discrete and Mixed RV Consider a discrete random variable X with range $R_X = \{x_1, \ldots, x_n\}$ and PMF $P_X(x_k)$. Note that the CDF for X can be written as

$$F_X(x) = \sum_{x_k \in R_X} P_X(x_k)u(x - x_k).$$

where:

- $u(x - x_k)$ is the Heaviside step function, which is defined as:

$$u(x - x_k) = \begin{cases} 0 & \text{if } x < x_k \\ 1 & \text{if } x \geq x_k \end{cases}$$

The sum $\sum_{x_k \in \mathbb{R}_X} P_X(x_k)u(x - x_k)$ effectively includes only those x_k values that are less than or equal to x due to the step function $u(x - x_k)$. Therefore, it accumulates the probabilities $P_X(x_k)$ for all $x_k \leq x$.

Now that we have symbolically defined the derivative of the step function as the delta function, we can write a PDF for X by "differentiating" the CDF:

$$\begin{aligned} f_X(x) &= \frac{dF_X(x)}{dx} \\ &= \sum_{x_k \in R_X} P_X(x_k)\frac{d}{dx}u(x - x_k) \\ &= \sum_{x_k \in R_X} P_X(x_k)\delta(x - x_k) \end{aligned}$$

We call this the **generalized PDF**.

$$EX = \int_{-\infty}^{\infty} x f_X(x)dx.$$

$$\begin{aligned} EX &= \int_{-\infty}^{\infty} x f_X(x) dx \\ &= \int_{-\infty}^{\infty} x \sum_{x_k \in R_X} P_X(x_k) \delta(x - x_k) dx \\ &= \sum_{x_k \in R_X} P_X(x_k) \int_{-\infty}^{\infty} x \delta(x - x_k) dx \\ &= \sum_{x_k \in R_X} x_k P_X(x_k) \end{aligned}$$

Chapter 5

Gaussian Distribution

5.1 Gaussian Distribution

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

We first define the standard normal random variable. We will then see that we can obtain other normal random variables by scaling and shifting a standard normal random variable.

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

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

The $1/\sqrt{2\pi}$ is there to make sure that the area under the PDF is equal to one.

More generally,

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \quad \text{for all } z \in \mathbb{R}.$$

5.1.1 Cumulative Distribution Function

The CDF of the standard normal distribution is denoted by the Φ function:

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

Here are some properties of the Φ function:

- $\lim_{x\to\infty} \Phi(x) = 1$

- $\lim_{x\to-\infty}\Phi(x)=0$
- $\Phi(0)=\frac{1}{2}$
- $\Phi(-x)=1-\Phi(x),\forall x\in\mathbb{R}$.

5.1.2 Multinomial

For a D-dimensional vector $\mathbf{x}$, the multivariate Gaussian distribution takes the form

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})=\frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \tag{5.1}$$

The multivariate Gaussian distribution is fully characterized by a mean vector and a covariance matrix.

5.1.3 Marginal and Conditional Gaussian Distributions

Consider first the case of conditional distributions. Suppose $\mathbf{x}$ is a D-dimensional vector with Gaussian distribution $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})$ and that we partition $\mathbf{x}$ into two disjoint subsets $\mathbf{x}_a$ and $\mathbf{x}_b$. Thus, $\mathbf{x}_a$ has M components and $\mathbf{x}_b$ has $D-M$ components.

$$\mathbf{x}=\begin{bmatrix}\mathbf{x}_a\\ \mathbf{x}_b\end{bmatrix}.$$

Similarly,

$$\boldsymbol{\mu}=\begin{bmatrix}\boldsymbol{\mu}_a\\ \boldsymbol{\mu}_b\end{bmatrix}$$

and the covariance matrix is given by

$$\boldsymbol{\Sigma}=\begin{bmatrix}\boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab}\\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb}\end{bmatrix}$$

Note that the symmetry $\boldsymbol{\Sigma}^T=\boldsymbol{\Sigma}$ implies that $\boldsymbol{\Sigma}_{ab}^T=\boldsymbol{\Sigma}_{ba}$.

$$p(\mathbf{x},\mathbf{y})=\mathcal{N}\left(\begin{bmatrix}\boldsymbol{\mu}_x\\ \boldsymbol{\mu}_y\end{bmatrix},\begin{bmatrix}\boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy}\\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy}\end{bmatrix}\right)$$

The conditional distribution $p(\mathbf{x}|\mathbf{y})$ is also Gaussian and given by

$$\begin{aligned}p(\mathbf{x}|\mathbf{y})&=\mathcal{N}(\boldsymbol{\mu}_{x|y},\boldsymbol{\Sigma}_{x|y})\\ \boldsymbol{\mu}_{x|y}&=\boldsymbol{\mu}_x+\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}(\mathbf{y}-\boldsymbol{\mu}_y)\\ \boldsymbol{\Sigma}_{x|y}&=\boldsymbol{\Sigma}_{xx}-\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx}.\end{aligned}$$

Note that the mean of the conditional distribution $p(\mathbf{x}|\mathbf{y})$ is a linear function of $\mathbf{y}$ and that the covariance, is independent of $\mathbf{x}$.

Note that the both the conditional and marginal covariance are equal only when $\boldsymbol{\Sigma}_{xy}=0$. This happens when $\mathbf{x}$ and $\mathbf{y}$ are uncorrelated.

The mean of the conditional distribution depends explicitly on new measurements ($\mathbf{y}$) whereas the conditional covariance does not depend on the new measurements. This means that we can know how our error will change once we receive our measurements, even before the measurements arrive.

5.1.4 Sampling from Multivariate Gaussian Distributions

To obtain samples from a multivariate normal $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we can use the properties of a linear transformation of a Gaussian random variable: If $\mathbf{x} \sim \mathcal{N}(0, I)$, then $\mathbf{y} = A\mathbf{x} + \boldsymbol{\mu}$, where $AA^T = \boldsymbol{\Sigma}$ is Gaussian distributed with mean and covariance matrix. One convenient choice of A is to use the Cholesky decomposition of the covariance matrix. The Cholesky decomposition has the benefit that A is triangular, leading to efficient computation:

$$Cov(\mathbf{y}) = Cov(A\mathbf{x}) = ACov(\mathbf{x})A^T = AIA^T = AA^T = \boldsymbol{\Sigma}.$$

Thus, we can use the Cholesky decomposition.

Chapter 6

Joint Distributions

In real life, we are often **interested in several random variables that are related to each other.** For example, suppose that we choose a random family, and we would like to study the number of people in the family, the household income, the ages of the family members, etc. Each of these is a random variable, and we suspect that they are dependent. In this chapter, we develop tools to study joint distributions of random variables. The concepts are similar to what we have seen so far. The only difference is that instead of one random variable, we consider two or more. In this chapter, we will focus on two random variables, but once you understand the theory for two random variables, the extension to n random variables is straightforward. We will first discuss joint distributions of discrete random variables and then extend the results to continuous random variables.

6.1 Joint PMF

Recall that for a discrete RV X, we define the PMF as $P_X(x) = P(X = x)$. Now, if we have two RVs X and Y, we define the joint PMF as follows:

$$P_{XY}(x, y) = P(X = x, Y = y).$$

Note that the comma means "and", so we can write as

$$\begin{aligned} P_{XY}(x, y) &= P(X = x, Y = y) \\ &= P(X = x \text{ and } Y = y) \\ &= P(X = x \cap Y = y) \end{aligned}$$

We can define the joint range for X and Y as

$$R_{XY} = \{(x, y) | P_{XY}(x, y) > 0\}.$$

In particular, if $R_X = \{x_1, x_2, \dots\}$ and $R_Y = \{y_1, y_2, \dots\}$, then

$$\begin{aligned} R_{XY} &\subset R_X \times R_Y \\ &= R_{XY} = \{(x_i, y_j) | x_i \in R_X, y_j \in R_Y\}. \end{aligned}$$

For two discrete RVs, we have

$$\sum_{(x_i, y_j) \in R_{XY}} P_{XY}(x_i, y_j) = 1$$

We can use the joint PMF to find $P((X,Y) \in A)$ for any set $A \subset \mathbb{R}^2$. Specifically, we have

$$P((X,Y) \in A) = \sum_{(x_i,y_j)\in(A\cap R_{XY})} P_{XY}(x_i, y_j)$$

6.2 Joint CDF

The joint cumulative distribution function of two random variables X and Y is defined as

$$F_{XY}(x,y) = P(X \leq x, Y \leq y).$$

Equivalently,

$$\begin{aligned} F_{XY}(x,y) &= P(X \leq x, Y \leq y) \\ &= P(X \leq x \cap Y \leq y) \end{aligned}$$

If we know the CDF of X and Y, we can find the *marginal* CDFs, $F_X(x)$ and $F_Y(y)$. Specifically, for any $x \in \mathbb{R}$, we have

$$\begin{aligned} F_{XY}(x,\infty) &= P(X \leq x, Y \leq \infty) \\ &= P(X \leq x) \\ &= F_X(x) \end{aligned}$$

6.3 Conditioning and Independence

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

6.3.1 Conditional PMF and CDF

The conditional PMF of X given an event A is given by

$$\begin{aligned} P_{X|A}(x_i) &= P(X = x_i|A) \\ &= \frac{P(X = x_i \text{ and } A)}{P(A)} \end{aligned}$$

Similarly,

$$F_{X|A}(x) = P(X \leq x|A)$$

6.3.2 Conditional PMF of X given Y

We have observed the value of a random variable Y, and we need to update the PMF of another random variable X whose value has not yet been observed. In these problems, we use the

conditional PMF of X given Y:

$$\begin{aligned} P_{X|Y}(x_i|y_j) &= P(X = x_i|Y = y_j) \\ &= \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)} \\ &= \frac{P_{XY}(x_i, y_j)}{P_Y(y_j)} \end{aligned}$$

6.3.3 Independent Random Variables

Two discrete RVs X and Y are independent if

$$P_{XY}(x, y) = P_X(x)P_Y(y), \forall x, y.$$

Equivalently,

$$F_{XY}(x, y) = F_X(x)F_Y(y), \forall x, y.$$

If X and Y are independent,

$$P_{X|Y}(x_i|y_j) = P_X(x_i).$$

6.3.4 Conditional Expectation

Given that we know an event A has occurred, we can compute the conditional expectation of a RV X, $E[X|A]$:

$$E[X|A] = \sum_{x_i \in R_X} x_i P_{X|A}(x_i).$$

Similarly, given that we have observed the value of random variable Y, we can compute the conditional expectation of X:

$$E[X|Y = y] = \sum_{x_i \in R_X} x_i P_{X|Y}(x_i|y).$$

6.4 The Law of Total Probability

Recall that the law of total probability: If $B_1, B_2, \dots$ is a partition of the sample space S, then for any event A we have

$$P(A) = \sum_i P(A \cap B_i) = \sum_i P(A|B_i)P(B_i).$$

If Y is a discrete random variable with range $R_y = \{y_1, y_2, \dots\}$, then the events $\{Y = y_1\}, \{Y = y_2\}, \dots$, form a partition of the sample space. Thus, we can use the law of total probability:

$$P_X(x) = \sum_{y_j \in R_Y} P_{XY}(x, y_j) = \sum_{y_j \in R_Y} P_{X|Y}(x|y_j)P_Y(y_j).$$

We can write this more generally as

$$P(X \in A) = \sum_{y_j \in R_Y} P(X \in A | Y = y_j) P_Y(y_j), \text{ for any set} A.$$

Similarly, we can write the law of total expectation:

$$EX = \sum_i E[X|B_i] P(B_i)$$
$$EX = \sum_{y_j \in R_Y} E[X|Y = y_i] P_Y(y_j).$$

This means that the expected value of X can be calculated from the probability distribution of $X|Y$ and Y, which is often useful both in theory and practice.

6.5 Functions of Two Random Variables

Suppose that you have two discrete random variables X and Y, and suppose that $Z = g(X, Y)$, where $g : \mathbb{R}^2 \to \mathbb{R}$. Then, the PMF of Z is given by

$$\begin{aligned} P_Z(z) &= P(g(X, Y) = z) \\ &= \sum_{(x_i, y_j) \in A_z} P_{XY}(x_i, y_j), \end{aligned}$$

where $A_z = \{(x_i, y_j) \in R_{XY} : g(x_i, y_j) = z\}$. Note that if we are only interested in $E[g(X, Y)]$, we can directly use LOTUS, without finding $P_Z(z)$:

$$E[g(X, Y)] = \sum_{(x_i, y_j) \in R_{XY}} g(x_i, y_j) P_{XY}(x_i, y_j).$$

6.6 Conditional Expectation and Conditional Variance

6.6.1 Conditional Expectation as a Function of a Random Variable

Note that

- $E[X]$ is a scalar value
- $E[X|Y]$ is a random variable, because the value depends on Y.

$$E[X] = \sum_x x \cdot p(x)$$
$$E_Y[E[X|Y]] = E[X]$$

Since, $E[E[X|Y]]$ is the function of Y. It is also called *the law of iterated expectations* or *the law of total expectation.*

$$\begin{aligned}
\mathbb{E}_Y[\mathbb{E}[X|Y]] &= \mathbb{E}_Y\left[\sum_x x \cdot P(X=x|Y)\right] \\
&= \sum_y \left[\sum_x x \cdot P(X=x|Y)\right] \cdot P(Y=y) \\
&= \sum_y \sum_x x \cdot P(X=x,Y) \\
&= \sum_x x \sum_y \cdot P(X=x,Y) \\
&= \sum_x x \cdot P(X=x) \\
&= \mathbb{E}[X]
\end{aligned}$$

To get an intuition, consider the following example: Let X be the lifetime duration of a lightbulb, and let Y be the factory the lightbulb was produced in. Suppose $E[X|Y=1] = 5000$ and $E[X|Y=2] = 4000$, indicating that factory 1 produces longer lasting bulbs. Suppose factory 1 supplies 60% of the lightbulbs, so $p(Y=1) = 0.6$ and $p(Y=2) = 0.4$. Then the expected duration of a random lightbulb is given by

$$E[X] = E[X|Y=1]p(Y=1) + E[X|Y=2]p(Y=2)$$

There is a similar formula for the variance. In particular, *the law of total variance*, also called the conditional variance formula, tells us that

$$V[X] = E_Y[V[X|Y]] + V_Y[E[X|Y]]$$

To see this, let's define the conditional moments, $\mu_{X|Y} = E[X|Y]$, $s_{X|Y} = E[X^2|Y]$, and $\sigma^2_{X|Y} = V[X|Y] = s_{X|Y} - \mu^2_{X|Y}$, which are functions of Y. Then, we have

$$\begin{aligned}
V[X] &= E[X^2] - (E[x])^2 = E_Y[s_{X|Y}] - (E_Y[\mu_{X|Y}])^2 \\
&= E_Y[\sigma^2_{X|Y}] + E_Y[\mu^2_{X|Y}] - (E_Y[\mu_{X|Y}])^2 \\
&= E_Y[V[X|Y]] + V_Y[\mu_{X|Y}]
\end{aligned}$$

$$
\begin{aligned}
E[Y \mid X = x] &= \sum_y y \cdot p_{Y|X}(y \mid X = x) \\
&= \sum_y y \cdot \frac{p_{X,Y}(x,y)}{p_X(x)} \\
&= \sum_y y \cdot \frac{\sum_z p_{X,Y,Z}(x,y,z)}{p_X(x)} \\
&= \sum_y y \cdot \frac{\sum_z p_{Y|X,Z}(y \mid X = x, Z = z) \cdot p_{X,Z}(x,z)}{p_X(x)} \\
&= \sum_z \frac{p_{X,Z}(x,z)}{p_X(x)} \sum_y y \cdot p_{Y|X,Z}(y \mid X = x, Z = z) \\
&= \sum_z p_{Z|X}(z \mid X = x) \cdot \sum_y y \cdot p_{Y|X,Z}(y \mid X = x, Z = z) \\
&= \sum_z p_{Z|X}(z \mid X = x) \cdot E[Y \mid X = x, Z = z) \\
&= E\left[E[Y \mid X, Z] \mid X = x\right]
\end{aligned}
$$

Note that if X and Y are independent,

- $E[X|Y] = EX$.
- $E[g(X)|Y] = E[g(X)]$.
- $E[XY] = EXEY$.
- $E[g(X)h(Y)] = E[g(X)]E[h(Y)]$.

6.6.2 Conditional Variance

We can define the conditional variance of X, $Var(X|Y = y)$. Let $\mu_{X|Y}(y) = E[X|Y = y]$, then

$$
\begin{aligned}
Var(X|Y = y) &= E[\big(X - \mu_{X|Y}(y)\big)^2 |Y = y] \\
&= \sum_{x_i \in R_X} \big(x_i - \mu_{X|Y}(y)\big)^2 P_{X|Y}(x_i) \\
&= E[X^2|Y = y] - \mu_{X|Y}(y)^2
\end{aligned}
$$

Note that $Var(X|Y = y)$ is a function of y.

6.7 Two Continuous Random Variables

6.7.1 Joint Probability Density Function

Two random variables are jointly continuous if they have a joint probability density function as follows:

Two random variables X and Y are jointly continuous if there exists a non-negative function $f_{XY} : \mathbb{R}^2 \to \mathbb{R}$, such that, for any set $A \in \mathbb{R}^2$, we have

$$P((X,Y) \in A) = \iint_A f_{XY}(x,y)dxdy.$$

The function $f_{XY}(x,y)$ is called the joint probability density function of X and Y. The domain of $f_{XY}(x,y)$ is the entire $\mathbb{R}^2$ and the range is

$$R_{XY} = \{(x,y)|f_{XY}(x,y) > 0\}.$$

The intuition behind the joint density is similar to that of the PDF of a single random variable. Recall that a random variable X and small positive δ, we have

$$P(x < X \leq x + \delta) \approx f_X(x)\delta.$$

Similarly, for small δ_x and δ_y,

$$P(x < X \leq x + \delta_x, y \leq Y \leq y + \delta_y) \approx f_{XY}(x,y)\delta_x\delta_y.$$

6.7.2 Joint CDF

6.7.3 Conditioning and Independence

6.7.4 Functions of Two Continuous Random Variables

LOTUS for two continuous random variables:

$$E[g(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f_{XY}(x,y)dxdy$$

The Method of Transformations

Theorem 1 *The Method of Transformations* Let X and Y be two jointly continuous random variables. Let $(Z,W) = g(X,Y) = (g_1(X,Y), g_2(X,Y))$, where $g : \mathbb{R}^2 \to \mathbb{R}^2$ is a continuous one-to-one (invertible) function with continuous partial derivatives. Let $h = g^{-1}$, *i.e.,* $(X,Y) = h(Z,W) = (h_1(Z,W), h_2(Z,W))$. Then Z and W are jointly continuous and their joint PDF, $f_{ZW}(z,w)$ for $(z,w) \in R_{ZW}$ is given by

$$f_{ZW}(z,w) = f_{XY}(h_1(z,w), h_2(z,w))|J|,$$

where J is the Jacobian of h defined by

$$\begin{aligned} J &= \det \begin{bmatrix} \frac{\partial h_1}{\partial z} & \frac{\partial h_1}{\partial w} \\ \frac{\partial h_2}{\partial z} & \frac{\partial h_2}{\partial w} \end{bmatrix} \\ &= \frac{\partial h_1}{\partial z} \cdot \frac{\partial h_2}{\partial w} - \frac{\partial h_2}{\partial z}\frac{\partial h_w}{\partial w}. \end{aligned}$$

Example Let X and Y be two independent standard normal RVs. Let

$$\begin{cases} Z &= 2X - Y \\ W &= -X + Y \end{cases}$$

Find $f_{ZW}(z, w)$.

X and Y are jointly continuous and their joint PDF is given by

$$f_{XY}(x, y) = f_X(x) f_Y(y)$$

Here, the function g is defined by $(z, w) = g(x, y) = (g_1(x, y), g_2(x, y)) = (2x - y, -x + y)$. We can obtain the inverse function h:

$$\begin{cases} x &= z + w = h_1(z, w) \\ y &= z + 2w = h_w(z, w) \end{cases}$$

Example Let X and Y be two RVs with joint PDF $f_{XY}(x, y)$. Let $Z = X + Y$. Find $f_Z(z)$.

To apply the above theorem, we need two random variables Z and W. We can simply define $W = X$. Then, we get

$$\begin{cases} z &= x + y \\ w &= x \end{cases}$$

Then, we can find the inverse transform:

$$\begin{cases} x &= w \\ y &= z - w \end{cases}$$

Thus,

$$f_{ZW}(z, w) = f_{XY}(w, z - w).$$

However, we are interested in the marginal PDF, $f_Z(z)$, we can get it by

$$f_Z(z) = \int_{-\infty}^{\infty} f_{XY}(w, z - w) dw.$$

Note that if X and Y are independent, then $f_{XY}(x, y) = f_X(x) f_Y(y)$ and we conclude that

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(w) f_Y(z - w) dw.$$

The above integral is called the *convolution* of f_X and f_Y, and we write

$$\begin{aligned} f_Z(z) &= f_X * f_Y \\ &= \int_{-\infty}^{\infty} f_X(w) f_Y(z - w) dw = \int_{-\infty}^{\infty} f_Y(w) f_X(z - w) dw \end{aligned}$$

The convolution can be thought as a sum of two independent RVs. For example, suppose we roll two dice. Let $y = x_1 + x_2$ be the sum of the dice. We have

$$\begin{aligned} p(y = 2) &= p(x_1 = 1) p(x_2 = 1) = \frac{1}{36} \\ p(y = 3) &= p(x_1 = 1) p(x_2 = 2) + p(x_1 = 2) p(x_2 = 1) = \frac{2}{36} \\ &\vdots \end{aligned}$$

Example Let X and Y be two independent RVs, where both are Gaussians and let $Z = X + Y$. Find the PDF of Z.

We have

$$\begin{aligned} f_Z(z) &= f_X(x) + f_Y(y) \\ &= \int_{-\infty}^{\infty} f_X(w) f_Y(z-w) dw \\ &= \int_{-\infty}^{\infty} \frac{1}{2\pi} e^{-\frac{w^2}{2}} e^{-\frac{(z-w)^2}{2}} dw \\ &= \frac{1}{\sqrt{4\pi}} e^{-\frac{z^2}{4}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi}} e^{-\left(w-\frac{z}{2}\right)^2} dw \\ &= \frac{1}{\sqrt{4\pi}} e^{-\frac{z^2}{4}} \end{aligned}$$

Note that if $X \sim \mathcal{N}\left(\mu_X, \sigma_X^2\right)$ and $Y \sim \mathcal{N}\left(\mu_Y, \sigma_Y^2\right)$, then

$$X + Y \sim \mathcal{N}\left(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2\right).$$

Hence, the convolution of two Gaussians is a Gaussian.

6.8 Covariance and Correlation

Consider two RVs X and Y. Here, we define the covariance between X and Y, $\text{Cov}(X, Y)$. The covariance gives some information about how X and Y are statistically related. The covariance between X and Y is defined as

$$\text{Cov}(X, Y) = E[(X - EX)(Y - EY)] = E[XY] - (EX)(EY)$$

Note that

$$\begin{aligned} E[(X - EX)(Y - EY)] &= E[XY - X(EY) - (EX)Y + (EX)(EY)] \\ &= E[XY] - (EX)(EY) - (EX)(EY) + (EX)(EY)] \\ &= E[XY] - (EX)(EY). \end{aligned}$$

Intuitively, the covariance between X and Y indicates how the values of X and Y move relative to each other. If large values of X tend to happen with large values of Y, then the covariance is positive and we say X and Y are positively correlated.

The covariance has the following properties:

- $\text{Cov}(X, X) = \text{Var}(X)$
- If X and Y are independent, then $\text{Cov}(X, Y) = 0$, since $E[XY] = EXEY$, so it is zero.
- $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
- $\text{Cov}(aX, Y) = a\text{Cov}(X, Y)$
- $\text{Cov}(X + c, Y) = \text{Cov}(X, Y)$

- $\text{Cov}(X+Y, Z) = \text{Cov}(X, Z) + \text{Cov}(Y, Z)$
- More generally,

$$\text{Cov}\left(\sum_{i=1}^{m} a_i X_i, \sum_{j=1}^{n} b_j Y_j\right) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j \text{Cov}(X_i, Y_j).$$

6.8.1 Variance of a Sum

$$\text{Var}(aX + bY) = a^2\text{Var}(X) + b^2\text{Var}(Y) + 2ab\text{Cov}(X, Y)$$

6.8.2 Correlation Coefficient

The *correlation coefficient*, denoted by ρ_{XY} or $\rho(X, Y)$ is obtained by normalizing the covariance. We can define the correlation coefficient of two random variables X and Y as the covariance of the standardized versions of X and Y,

$$\begin{aligned}\rho_{XY} &= \text{Cov}(\frac{X - EX}{\sigma_X}, \frac{Y - EY}{\sigma_Y}) \\ &= \text{Cov}(\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y}) \quad \text{by the property of Cov} \\ &= \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}\end{aligned}$$

A nice property of the correlation coefficient is that it is always between -1 and 1. This is an immediate result of *Cauchy-Schwarz inequality*. One way to prove that $-1 \leq \rho \leq 1$ is to use the following inequality:

$$\alpha\beta \leq \frac{\alpha^2 + \beta^2}{2}, \quad \text{for } \alpha, \beta \in \mathbb{R}.$$

This is because $(\alpha - \beta)^2 \geq 0$. The equality holds only when $\alpha = \beta$. From this, we can conclude that for any two random variables U and V, which are the standardized versions of X and Y, respectively:

$$E[UV] \leq \frac{EU^2 + EV^2}{2}.$$

By the definition, $\rho_{XY} = \text{Cov}(U, V) = E[UV]$. Note that $EU^2 = EV^2 = 1$ by definition, so we get

$$\rho_{XY} = E[UV] \leq \frac{EU^2 + EV^2}{2} = 1,$$

with equality only if $U = V$.

Note that two independent random variables are always uncorrelated, but the converse is not necessary true. In other words, if X and Y are uncorrelated, then X and Y may or may not be independent.

6.9 Bivariate Normal Distribution

6.9.1 Mixed Case

The mixed joint density may be defined where one or more random variables are continuous and the other random variables are discrete. With one variable of each type

$$f_{XY}(x, y) = f_{X|Y}(x|y)P_Y(Y = y) = P(Y = y|X = x)f_X(x)$$

Chapter 7

Multiple Random Variables

7.1 Joint Distributions and Independence

For three or more random variables, the joint PDF, joint PMF, and joint CDF are defined in a similar way to wat we have already seen for the case of two random variables. Let $X_1, \ldots, Xn$ be n discrete random variables. The joint PMF of $X_1, \ldots, X_n$ is defined as

$$P_{X_1,\ldots X_n}(x_1, \ldots, x_n) = P(X_1 = x_1, \ldots, X_n = x_n).$$

For n jointly continuous random variables $X_1, \ldots, Xn$ the joint PDF is defined to be the function $f_{X_1,\ldots,Xn}(x_1, \ldots, x_n)$ such that the probability of any set $A \subset \mathbb{R}^n$ is given by the integral of the PDF over the set A. In particular, for a set $A \in \mathbb{R}^n$, we can write

$$P\left((X_1, \ldots X_n) \in A\right) \int \cdots \int_A \cdots \int f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) dx_1, \ldots, dx_n.$$

The marginal PDF of X_i can be obtained by integrating all other X_j's. For example,

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) dx_1, \ldots, dx_n.$$

The joint CDF of n random variables $X_1, \ldots, X_n$ is defined as

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = P(X_1 \leq x_1, \ldots, X_n \leq x_n).$$

Independence The idea of Independence is exactly the same as what we have seen before:

- $F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = F_{X_1}(x_1) F_{X_2}(x_2) \cdots F_{X_n}(x_n)$.
- Equivalently, if $X_1, \ldots, X_n$ are discrete, then they are independent if for all

$$P_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = P_{X_1}(x_1) P_{X_2}(x_2) \cdots P_{X_n}(x_n).$$

- If $X_1, \ldots, X_n$ are continuous, then they are independent if for all

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = f_{X_1}(x_1) f_{X_2}(x_2) \cdots f_{X_n}(x_n).$$

- If random variables are independent,

$$E[X_1, \ldots, X_n] = E[X_1] \ldots E[X_n].$$

If random variables $X_1, \ldots, X_n$ are independent and identically distributed (i.i.d.) then they will have the same means and variances, so we can write

$$\begin{aligned} E[X_1, \ldots, X_n] &= E[X_1] \ldots E[X_n] \quad \text{since, the they are independent} \\ &= E[X_1] \ldots E[X_1] \quad \text{since, the they are identically distributed} \\ &= E[X_1]^n \end{aligned}$$

7.2 Sums of Random Variables

A random variable Y is given by

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

The linearity of expectations tells us that

$$EY = EX_1 + \cdots + EX_n.$$

We can also find the variance of Y.

$$\text{Var}(X_1 + X_2) = \text{Var}(X_1) + \text{Var}(X_2) + 2\text{Cov}(X_1, X_2).$$

For $Y = X_1 + \cdots + X_n$, we can obtain a more general version of the above equation.

$$\begin{aligned} \text{Var}(X_1 + X_2) &= \text{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j\right) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} \text{Cov}(X_i, X_j) \\ &= \sum_{i=1}^{n} \text{Var}(X_i) + 2 \sum_{i<j}^{n} \text{Cov}(X_i, X_j). \end{aligned}$$

If the X_i's are independent, then $\text{Cov}(X_i, X_j) = 0$ for $i \neq j$.

7.3 Moment Generating Functions

The n-th moment of a random variable X is defined to be $E[X^n]$. The n-th central moment of X is defined to be $E[(X - EX)^n]$.

For instance, the first moment is the expected value $E[X]$. The second central moment is the variance of X. The moment generating function (MGF) of a random variable X is a function $M_X(s)$ defined as

$$M_X(s) = E[e^{sX}].$$

We say that MGF of X exists, if there exists a positive constant α such that $M_X(s)$ is finite for all $s \in [-a, a]$.

7.3.1 Sum of Independent Random Variables

Suppose $X_1, \ldots, X_n$ are n independent random variables, and the random variable Y is defined as

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

Then ,

$$\begin{aligned} M_Y(s) &= E[e^{sY}] \\ &= E[e^{s(X_1+\cdots+X_n)}] \\ &= E[e^{sX_1}e^{sX_2}\ldots e^{sX_n}] \\ &= E[e^{sX_1}]\ldots E[e^{sX_n}] \quad \text{since, they are independent} \\ &= M_{X_1}(s)M_{X_2}(s)\ldots M_{X_n}(s) \end{aligned}$$

7.4 Characteristic Functions

There are some random variables for which the moment generating function does not exist on any real interval with positive length. In that case, we can use the characteristic function defined as

$$\phi_X(\omega) = E[e^{j\omega X}],$$

where $j = \sqrt{-1}$ and ω is a real number. Note that if X is a real-valued random variable, we can write $|e^{j\omega X}| = 1$. Therefore, we conclude

$$\begin{aligned} |\phi_X(\omega)| &= |E\left[e^{j\omega X}\right]| \\ &\leq |E\left[e^{j\omega X}\right]| \\ &\leq 1 \end{aligned}$$

The characteristic function has similar properties to the MGF. If $X_1, \ldots, X_n$ are n independent random variables, then

$$\phi_{X_1+\cdots+X_n}(\omega) = \phi_{X_1}(\omega)\ldots\phi_{X_n}(\omega).$$

7.5 Random Vectors

When we have n random variables, we can put them in a vector $\mathbf{X}$:

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$$

We call $\mathbf{X}$ a n-dimensional random vector.

For a random vector $\mathbf{X}$, we defined the **correlation matrix**, $\mathbf{R_X}$, as

$$\mathbf{R_X} = E[\mathbf{X}\mathbf{X}^T] = \begin{bmatrix} X_1^2 & X_1X_2 & \ldots & X_1X_n \\ \vdots & \ddots & \vdots & \vdots \\ X_nX_1 & X_nX_2 & \ldots & X_n^2 \end{bmatrix} = \begin{bmatrix} EX_1^2 & E[X_1X_2] & \ldots & E[X_1X_n] \\ \vdots & \ddots & \vdots & \vdots \\ E[X_nX_1] & E[X_nX_2] & \ldots & E[X_n^2] \end{bmatrix}$$

The covariance matrix, $\mathbf{C}_X$, is defined as

$$\begin{aligned}\mathbf{C_X} &= E[(\mathbf{X} - E\mathbf{X})(\mathbf{X}^T - E\mathbf{X})^T] \\ &= \begin{bmatrix} (X_1 - EX_1)^2 & (X_1 - EX_1)(X_2 - EX_2) & \dots & (X_1 - EX_1)(X_n - EX_n) \\ \vdots & \ddots & \vdots & \vdots \\ (X_n - EX_n)(X_1 - EX_1) & (X_n - EX_n)(X_2 - EX_2) & \dots & (X_n - EX_n)^2 \end{bmatrix} \\ &= \begin{bmatrix} \mathrm{Var}(X_1)^2 & \mathrm{Cov}(X_1, X_2) & \dots & \mathrm{Cov}(X_1, X_n) \\ \vdots & \ddots & \vdots & \vdots \\ \mathrm{Cov}(X_n, X_1) & \mathrm{Cov}(X_n, X_2) & \dots & \mathrm{Var}(X_n) \end{bmatrix}\end{aligned}$$

The covariance matrix is a generalization of the variance of a random variable.

Let $\mathbf{X}$ be an n-dimensional random vector and the random vector $\mathbf{Y}$ be defined as

$$\mathbf{Y} = \mathbf{AX} + \mathbf{b},$$

where $\mathbf{A}$ is a fixed $m \times n$ matrix and $\mathbf{b}$ is a fixed m-dimensional vector. Then,

$$\mathbf{C_Y} = \mathbf{AC_X A^T}.$$

7.5.1 Properties of the Covariance Matrix

The covariance matrix is the generalization of the variance to random vectors. It is an important matrix and is used extensively. Let's take a moment and discuss its properties. Here, we use concepts from linear algebra such as eigenvalues and positive definiteness. First note that, for any random vector $\mathbf{X}$, the covariance matrix $\mathbf{C_X}$ is a **symmetric matrix**. This is because if $\mathbf{C_X} = [c_{ij}]$, then

$$c_{ij} = \mathrm{Cov}(X_i, X_j) = \mathrm{Cov}(X_j, X_i) = c_{ji}.$$

Thus, the covariance matrix has all the nice properties of symmetric matrices. In particular, $\mathbf{C_X}$ can be diagonalized and all the eigenvalues of $\mathbf{C_X}$ are real. Here, we assume $\mathbf{X}$ is a real random vector. *i.e.,* the X_i can only take real values. A special property of the covariance matrix is that it is positive semi-definite (PSD). A symmetric matrix $\mathbf{M}$ is PSD if

$$\mathbf{b^T M b} \geq 0.$$

To show that $\mathbf{C_X}$ is always PSD, let $\mathbf{b}$ be any fixed vector with n elements. Define the random variable Y as

$$Y = \mathbf{b^T}(\mathbf{X} - \mathbf{EX}).$$

We have

$$\begin{aligned} 0 &\leq EY^2 \\ &= E(YY^T) \\ &= \mathbf{b^T} E[(\mathbf{X} - \mathbf{EX})(\mathbf{X} - \mathbf{EX})^T]\mathbf{b} \\ &= \mathbf{b^T C_X b} \end{aligned}$$

Note that the eigenvalues of a PSD matrix are always larger than or equal to zero. If all the eigenvalues are strictly larger than zero, then the matrix is positive definite. From linear algebra, we know that a real symmetric matrix is positive definite if and only if all its eigenvalues are positive.

7.5.2 Functions of Random Vectors: The Method of Transformations

A function of a random vector is a random vector. Let $\mathbf{X}$ be an n-dimensional random vector with joint PDF $f_{\mathbf{X}}\mathbf{x}$ and $G : \mathbb{R}^n \to \mathbb{R}^n$ be a continuous and invertible function with continuous partial derivatives and let $H = G^{-1}$. Suppose that the random vector $\mathbf{Y}$ is given by $\mathbf{Y} = G(\mathbf{X})$ and thus $\mathbf{X} = G^{-1}(\mathbf{Y}) = H(\mathbf{Y})$. That is,

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} H_1(Y_1, \ldots, Y_n) \\ \vdots \\ H_n(Y_1, \ldots, Y_n) \end{bmatrix}$$

Then, the PDF of $\mathbf{Y}$ is $f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n)$, is given by

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}(H(\mathbf{y}))|J|,$$

where $|J|$ is the Jacobian of H,

$$J = \det \begin{bmatrix} \frac{\partial H_1}{\partial y_1} & \cdots & \frac{\partial H_1}{\partial y_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial H_n}{\partial y_1} & \cdots & \frac{\partial H_n}{\partial y_n} \end{bmatrix}$$

Let $\mathbf{X}$ be an n-dimensional random vector. Let $\mathbf{A}$ be a fixed invertible $n \times n$ matrix, and $\mathbf{b}$ be a fixed n-dimensional vector. A random vector $\mathbf{Y}$ is given by

$$\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{b}.$$

The PDF of $\mathbf{Y}$ can be obtained as follows:

$$\mathbf{X} = \mathbf{A}^{-\mathbf{1}}(\mathbf{Y} - \mathbf{b}).$$

$$J = \det(\mathbf{A}^{-1}) = \frac{1}{\det(\mathbf{A})}.$$

Thus,

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{|\det(\mathbf{A})|} f_{\mathbf{X}}(\mathbf{A}^{-1}(\mathbf{y} - \mathbf{b}))$$

7.6 Probability Bounds

7.6.1 The Union Bound and Extension

The **union bound** or **Boole's inequality** is applicable when you need to show that the probability of union of some events is less than some value. For any two events A and B we have

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &\leq P(A) + P(B) \end{aligned}$$

In general, for any events, $A_1, \ldots, A_n$, we have

$$P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i).$$

7.6.2 Markov Inequality

Let X be any positive continuous random variable, we can write

$$\begin{aligned} EX &= \int_{-\infty}^{\infty} x f_X(x) dx \\ &= \int_{0}^{\infty} x f_X(x) dx \quad \text{since } X \text{ is positive-valued} \\ &\geq \int_{a}^{\infty} x f_X(x) dx \\ &\geq \int_{a}^{\infty} a f_X(x) dx \\ &= a \int_{a}^{\infty} f_X(x) dx \\ &= aP(X \geq a). \end{aligned}$$

Thus, we conclude

$$P(X \geq a) \leq \frac{EX}{a}, \quad \text{for any } a > 0.$$

We can prove the above inequality for discrete or mixed random variables similarly (using the generalized PDF), so we have the following result, called *Markov's inequality*.

If X is any non-negative random variable, then

$$P(X \geq a) \leq \frac{EX}{a}, \quad \text{for any } a > 0.$$

7.6.3 Chebyshev's Inequality

Let X be any random variable. If you define $Y = (X - EX)^2$, then Y is a non-negative random variable, so we can apply Markov's inequality to Y. In particular, for any positive real number b, we have

$$P(X \geq b^2) \leq \frac{EX}{b^2}.$$

Note that

$$\begin{aligned} &EY = E(X - EX)^2 = \text{Var}(X), \\ &P(Y \geq b^2) = P\left((X - EX)^2 \geq b^2\right) = P(|X - EX| \geq b). \end{aligned}$$

Thus, we get

$$P(|X - EX| \geq b) \leq \frac{\text{Var}(X)}{b^2}.$$

7.6.4 Chernoff Bounds

If X is a random variable, then for any $a \in \mathbb{R}$, we can write

$$\begin{aligned} P(X \geq a) &= P(e^{sX} \geq e^{sa}), \quad \text{for } s > 0, \\ P(X \leq a) &= P(e^{sX} \geq e^{sa}), \quad \text{for } s < 0. \end{aligned}$$

Note that e^{sX} is always a positive random variable for all $s \in \mathbb{R}$. Thus, we can apply Markov's inequality. For $s > 0$, we can write

$$\begin{aligned} P(X \geq a) &= P(e^{sX} \geq e^{sa}) \\ &\leq \frac{E[e^{sX}]}{e^{sa}}. \end{aligned}$$

Similarly, for $s < 0$, we can write

$$\begin{aligned} P(X \leq a) &= P(e^{sX} \geq e^{sa}) \\ &\leq \frac{E[e^{sX}]}{e^{sa}}. \end{aligned}$$

Also note that $E[e^{sX}]$ is the moment generating function, $M_X(s)$. Thus, we conclude

$$\begin{aligned} P(X \geq a) &\leq e^{-sa} M_X(s), \quad \text{for all } s > 0, \\ P(X \leq a) &\leq e^{-sa} M_X(s), \quad \text{for all } s < 0. \end{aligned}$$

Since Chernoff bounds are valid for all values of $s > 0$ and $s < 0$, we can choose s in a way to obtain the best bound, that is

$$\begin{aligned} P(X \geq a) &\leq \min_{s>0} e^{-sa} M_X(s) \\ P(X \leq a) &\leq \min_{s<0} e^{-sa} M_X(s) \end{aligned}$$

Comparison between Markov, Chebyshev, and Chernoff Bounds: For a random varible $X \sim \text{Binom}(n, p)$, upper bounds of $P(X \geq)$ of each bound when $p = \frac{1}{4}$ and $\alpha = \frac{3}{4}$ is given by

$$\begin{aligned} P(X \geq \frac{3n}{4}) &\leq \frac{2}{3}, \quad \text{Markov} \\ P(X \geq \frac{3n}{4}) &\leq \frac{4}{n}, \quad \text{Chebyshev} \\ P(X \geq \frac{3n}{4}) &\leq \frac{16}{27}^{\frac{n}{4}}, \quad \text{Chernoff.} \end{aligned}$$

The bound given by Markov is the weakest one. It is constant and does not change as n increases. The bound given by Chebyshev's inequality is stronger than the one given by Markov's inequality. The strongest bound is the Chernoff bound since it goes to zero exponentially.

7.6.5 Cauchy-Schwarz Inequality

For any two random variables X and Y, we have

$$|EXY| \leq \sqrt{E[X^2]E[Y^2]},$$

where equality holds if and only if $X = \alpha Y$, for some constant $\alpha \in \mathbb{R}$.

7.6.6 Jensen's Inequality

Chapter 8

Limit Theorems and Convergence of Random Variables

8.1 Law of Large Numbers

The law of large numbers has a very central role in probability and statistics. It states that **if you repeat an experiment independently a large number of times and average the result, what you obtain should be close to the expected value.** There are two main versions of the law of large numbers. They are called the *weak and strong laws of the large numbers*.

For i.i.d. random variables $X_1, \ldots, X_n$, the sample mean, denoted by $\overline{X}$, is defined as

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n}.$$

Another common notation for the sample mean is M_n. If the X_i's have CDF $F_X(x)$, we might show the sample mean by $M_n(X)$ to indicate distribution of the X_is.

Note that since the X_is are random variables, the sample mean, $\overline{X} = M_n(X)$, is also a random variable. In particular we have

$$\begin{aligned} E\overline{X} &= \frac{EX_1 + \cdots + EX_n}{n} \quad \text{by linearity of expectation} \\ &= \frac{nEX}{n} \quad \text{Since they are i.i.d., } EX_i = EX \\ &= EX. \end{aligned}$$

Also the variance of $\overline{X}$ is given by

$$\begin{aligned} \mathrm{Var}(\overline{X}) &= \frac{\mathrm{Var}(X_1 + \cdots + X_n)}{n^2} \quad \text{Since } \mathrm{Var}(aX)=a^2\mathrm{Var}(X) \\ &= \frac{\mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n)}{n^2} \quad \text{Since } X_i\text{s are independent} \\ &= \frac{n\mathrm{Var}(X)}{n^2} \\ &= \frac{\mathrm{Var}(X)}{n}. \end{aligned}$$

The weak law of large numbers (WLLN) states that for any $\epsilon > 0$, i.i.d. random variables $X_1, \ldots, X_n$ with a finite expected value $EX_i = \mu < \infty$,

$$\lim_{n\to\infty} P(|\overline{X} - \mu| \geq \epsilon) = 0.$$

8.2 Central Limit Theorems

The central limit theorem (CLT) is one of the most important results in probability theory. It states that, **under certain conditions, the sum of a large number of random variables is approximately normal**.

Suppose that $X_1, \ldots, X_n$ are i.i.d. random variables with expected values $EX_i = \mu < \infty$ and variance $\text{Var}(X_i) = \sigma^2 < \infty$.

Then as we saw above, the sample mean and variance have $E\overline{X} = \mu$ and $\text{Var}(\overline{X}) = \frac{\sigma^2}{n}$. Thus, the normalized random variable

$$Z_n = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{X_1 + \cdots + X_n - n\mu}{\sqrt{n}\sigma}$$

has zero mean $EZ_n = 0$ and variance $\text{Var}(Z_n) = 1$. The central limit theorem states that the CDF of Z_n converges to the standard normal CDF as n goes to infinity, that is

$$\lim_{n\to\infty} P(Z \leq x)\Phi(x), \quad \text{for all } x \in \mathbb{R},$$

where $\Phi(x)$ is the standard normal CDF.

In sum CLT states that **the CDF of Z_n is converging to the CDF of $\mathcal{N}(0, 1)$.**

The importance of the central limit theorem stems from the fact that, in many real applications, a certain random variable of interest is a sum of a large number of independent random variables. In these situations, we are often able to use the CLT to justify using the normal distribution. Examples of such random variables are found in almost every discipline. Here are a few:

- Laboratory measurement errors are usually modeled by normal random variables.
- In communication and signal processing, Gaussian noise is the most frequently used model for noise.
- In finance, the percentage changes in the prices of some assets are sometimes modeled by normal random variables.
- When we do random sampling from a population to obtain statistical knowledge about the population, we often model the resulting quantity as a normal random variable.

8.3 Convergence of Random Variables

We would like to see if a sequence of a random variables $X_1, X_2 \ldots$ converges to a random variable X. That is, we would like to see if X_n gets closer and closer to X as n increases.

In fact, we have already seen the concept of convergence when we discussed limit theorems (the weak law of large numbers (WLLN) and the central limit theorem (CLT)). The WLLN states that the average of a large number of i.i.d. random variables converges in probability to the expected value. The CLT states that the normalized average of a sequence of i.i.d. random variables converges in distribution to a standard normal distribution.

8.3.1 Convergence of a Sequence of Numbers

Let's say we have a sequence of a real numbers $a_1, a_2, \ldots$, which is defined as $a_n = \frac{n}{n+1}$. Then, we can ask whether the sequence converges. This sequence converges to 1. We say that a sequence converges to a limit L if a_n approaches to L as n goes to infinity. That is, for any $\epsilon > 0$, there exists an $N \in \mathbb{N}$ such that

$$|a_n - L| < \epsilon, \quad \text{for all } n > N.$$

8.3.2 Sequence of Random Variables

In any probability model, we have a sample space $S = \{s_1, \ldots, s_k\}$ and a probability measure P. Then, a random variable X is a mapping that assigns a real number to any of the possible outcomes s_i, $i = 1, 2, \ldots, k$. Thus, we may write

$$X(s_i) = x_i, \quad \text{for } i = 1, 2, \ldots, k.$$

8.3.3 Different Types of Convergence for Sequences of Random Variables

Consider a sequence of random variables $X_1, X_2, \ldots,$. This sequence might converge to a random variable X. There are four types of convergence that we are going to discuss:

- Convergence in distribution
- Convergence in probability
- Convergence in mean
- Almost sure convergence

Some of these convergence types are "stronger" than others and some are "weaker." By this, we mean the following: If Type A convergence is stronger than Type B convergence, it means that Type A convergence implies Type B convergence.

Convergence in Distribution Convergence in distribution is the weakest type of convergence in some sense. All it says is that the CDF of X_n's converges to the CDF of X as n goes to infinity. To say that X_n converges in distribution to X, we write

$$X_n \xrightarrow{d} X.$$

Convergence in Probability Convergence in probability is stronger than convergence in distribution. In particular, for a sequence $X_1, X_2, \ldots$ to converge to a random variable X, we must have that $P(|X_n - X| \geq \epsilon)$ goes to 0 as $n \to \infty$, for any $\epsilon > 0$.

$$X_n \xrightarrow{p} X.$$

Convergence in Mean One way of interpreting the convergence of a sequence X_n to X is to say that the "distance" between X and X_n is getting smaller and smaller. For instance, if we define the distance between X_n and X as $P(|X_n - X| \geq \epsilon)$, we have convergence in probability. One way to define the distance between X_n and X is

$$E(|X_n - X|^r),$$

where $r \geq 1$ is a fixed number. This refers to convergence in mean. The most common choice for r is 2, in which case it is called the *mean square convergence*.

Almost Sure Convergence

Chapter 9

Statistical Inference: Classical Methods

In real life, we work with data that are affected by randomness, and we need to extract information and draw conclusions from the data. The randomness might come from a variety of sources.

Statistical inference is a collection of methods that deal with drawing conclusions from data that are prone to random variation.

Frequentist (classical) Inference: In this approach, the unknown quantity θ is assumed to be a **fixed quantity**. That is, θ is a deterministic (non-random) quantity that is to be estimated by the observed data. For example, in the polling problem stated above we might consider θ as the percentage of people who will vote for a certain candidate, call him/her Candidate A. After asking n randomly chosen voters, we might estimate θ by

$$\theta = \frac{Y}{n},$$

where Y is the number of people who support for candidate A.

Bayesian Inference: In the Bayesian approach the unknown quantity θ is assumed to be a **random variable**, and we assume that we have some initial guess about the distribution of θ. After observing the data, we update the distribution of θ using Bayes' Rule.

9.1 Point Estimation

Here, we assume that θ is an unknown parameter to be estimated. For example, θ might be the expected value of a random variable, $\theta = EX$. The important assumption here is that θ is a fixed (non-random) quantity. To estimate θ, we need to collect some data. Specifically, we get a random sample $X_1, \ldots, X_n$ such that X_i's have the same distribution as X. To estimate θ, we define a point estimator $\hat{\theta}$ that is a function of the random sample, i.e.,

$$\hat{\theta} = h(X_1, \ldots, X_n).$$

For instances, if $\theta = EX$, we may choose $\hat{\theta}$ to be the sample mean,

$$\hat{\theta} = \overline{X} = \frac{X_1 + \cdots + X_n}{n}.$$

There are infinitely many possible estimators for θ, so how can we make sure that we have chosen a good estimator? How do we compare different possible estimators? To do this, we provide a

list of some desirable properties that we would like our estimators to have. Intuitively, we know that a good estimator should be able to give us values that are "close" to the real value of θ. To make this notion more precise we provide some definitions.

9.1.1 Evaluating Estimators

We define three main desirable properties for point estimators. The first one is related to **the estimator's bias**. The bias of an estimator $\hat{\theta}$ tells us on average how far $\hat{\theta}$ is from the real value of θ.

Let $\hat{\theta} = h(X_1, \ldots, X_n)$ be a point estimator for θ (*e.g.,* population mean). The bias of point estimator $\hat{\theta}$ is defined by

$$B(\hat{\theta}) = E[\hat{\theta}] - \theta.$$

In general, we would like to have a bias that is close to 0, indicating that on average, $\hat{\theta}$ is close to θ. It is worth noting that $B(\hat{\theta})$ might depend on the actual value of θ. In other words, you might have an estimator for which $B(\hat{\theta})$ is small for some values of θ and large for some other values of θ. A desirable scenario is when $B(\hat{\theta}) = 0$, i.e, $E[\hat{\theta}] = \theta$, for all values of θ. In this case, we say that $\hat{\theta}$ is an *unbiased estimator* of θ.

Let $\hat{\theta} = h(X_1, \ldots, X_n)$ be a point estimator for θ.We say that $\hat{\theta}$ is an unbiased estimator of θ if

$$B(\hat{\theta}) = 0, \text{ for all possible values of } \theta.$$

Note that if an estimator is unbiased, it does not necessarily means a good estimator.

Thus, we use another measure called *the mean squared error* which is given by

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

Example: Let $X_1, \ldots, X_n$ be a random sample from a distribution with $EX_i = \theta$, and variance $\text{Var}(X_i) = \sigma^2$. Consider the following two estimators:

- $\hat{\theta}_1 = X_1$
- $\hat{\theta}_2 = \overline{X_1} = \frac{X_1 + \cdots + X_n}{n}$

Find MSE of them and show that for $n > 1$.

1.

$$\begin{aligned} MSE(\hat{\theta}_1) &= E[(\hat{\theta}_1 - \theta)^2] \\ &= E[(X_1 - EX_1)^2] \\ &= \text{Var}(X_1) \\ &= \sigma^2 \end{aligned}$$

2.

$$\begin{aligned} MSE(\hat{\theta}_2) &= E[(\hat{\theta}_2 - \theta)^2] \\ &= E[(\overline{X} - \theta)^2] \\ &= \text{Var}(\overline{X} - \theta) + E[\overline{X} - \theta)^2]^2 \\ &\vdots \\ &= \sigma^2/n \end{aligned}$$

We leverage the equality from $EY^2 = \text{Var}(Y) + (EY)^2$, where $Y = \overline{X} - \theta$. Also, $\text{Var}(\overline{X} - \theta) = \text{Var}(\overline{X})$, since θ is a constant. Finally, $E[\overline{X} - \theta)^2] = 0$. Thus, we get the final result.

From the above example, we conclude that the both estimators are unbiased estimators of the mean, but the second one is better, since it has a smaller MSE. We can rewrite the second estimator as follows:

$$\begin{aligned} MSE(\hat{\theta}) &= E[(\hat{\theta} - \theta)^2] \\ &= \text{Var}(\hat{\theta} - \theta) + (E[\hat{\theta} - \theta])^2 \\ &= \underbrace{\text{Var}(\hat{\theta})}_{\text{Variance}} + \underbrace{B(\hat{\theta})^2}_{\text{Bias}} \end{aligned}$$

The last property that we discuss for point estimators is *consistency*. Loosely speaking, we say that an estimator is consistent if as the sample size n gets larger, $\hat{\theta}$ converges to the real value of θ. More precisely, we have the following definition:

Let $\hat{\theta} = h(X_1, \ldots, X_n)$ be a point estimator for θ. We say $\hat{\theta}_n$ is a *consistent estimator* of θ if

$$\lim_{n\to\infty} P(|\hat{\theta} - \theta| \geq \epsilon) = 0, \quad \text{for all } \epsilon > 0.$$

Example: Let $X_1, \ldots, X_n$ be a random sample with $EX_i = \theta$, and variance $\text{Var}(X_i) = \sigma^2$. Show that $\hat{\theta}_n = \overline{X}$ is a consistent estimator of θ: By the weak law of large numbers, we can apply Chebyshev's inequality to write

$$\begin{aligned} P(|\overline{X} - \theta| \geq \epsilon) &\leq \frac{\text{Var}(\overline{X})}{\epsilon^2} \\ &= \frac{\sigma^2}{n\epsilon^2} \end{aligned}$$

This goes to zero as $n \to \infty$. We already found that MSE goes to zero as $n \to \infty$. Therefore, it is a consistent estimator.

9.1.2 Point Estimators for Mean and Variance

The sample mean is an unbiased estimator of the mean.

$$\mathbb{E}[\bar{X}] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[X_i] = \mathbb{E}[X]$$

Suppose that we would like to estimate the variance of a distribution σ^2. Assuming $0 < \sigma^2 < \infty$, by definition

$$\sigma^2 = E[(X - \mu)^2].$$

Thus, the variance itself is the mean of the random variable $Y = (X - \mu)^2$. This suggests the following estimator for the variance

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \mu)^2.$$

By the linearity of expectation, $\hat{\sigma}^2$ is an unbiased estimator of σ^2. Also, by the weak law of large numbers, it is also a consistent estimator of σ^2. However, we often do not know about the true value of μ. Thus, we may replace μ by our estimate of the μ, the sample mean, to obtain the following estimator for σ^2.

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

To rearrange the above equation, we need some tricks. Suppose $X_1, \ldots, X_n$ are i.i.d. random variables with expectation μ and variance σ^2

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

Then S^2 is a biased estimator of σ^2, because

$$\begin{aligned}
\mathbb{E}[S^2] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\Big((X_i - \mu) - (\bar{X} - \mu)\Big)^2\right] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\Big((X_i - \mu)^2 - 2(X_i - \mu)(\bar{X} - \mu) + (\bar{X} - \mu)^2\Big)\right] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - \frac{2}{n}(\bar{X} - \mu)\sum_{i=1}^{n}(X_i - \mu) + \frac{1}{n}(\bar{X} - \mu)^2\sum_{i=1}^{n}1\right] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - \frac{2}{n}(\bar{X} - \mu)\sum_{i=1}^{n}(X_i - \mu) + (\bar{X} - \mu)^2\right]
\end{aligned}$$

To continue, we note that by subtracting μ from both sides of $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, we get

$$\bar{X} - \mu = \frac{1}{n}\sum_{i=1}^{n} X_i - \mu = \frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)$$

Then, the previous equation becomes:

$$\begin{aligned}
\mathbb{E}[S^2] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - \frac{2}{n}(\bar{X}-\mu)\sum_{i=1}^{n}(X_i-\mu) + (\bar{X}-\mu)^2\right] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - \frac{2}{n}(\bar{X}-\mu)\cdot n\cdot(\bar{X}-\mu) + (\bar{X}-\mu)^2\right] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - 2(\bar{X}-\mu)^2 + (\bar{X}-\mu)^2\right] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - (\bar{X}-\mu)^2\right] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2\right] - \mathbb{E}\Big[(\bar{X}-\mu)^2\Big] \\
&= \frac{1}{n}\mathbb{E}[n\cdot(X_i-\mu)^2] - \mathbb{E}\Big[(\bar{X}-\mu)^2\Big] = \sigma^2 - Var[\bar{X}] \\
&= \sigma^2 - Var\left[\frac{1}{n}\sum_{i=1}^{n}X_i\right] \\
&= \sigma^2 - \frac{1}{n^2}\sum_{i=1}^{n}Var(X_i) \\
&= \sigma^2 - \frac{1}{n^2}n\sigma^2 \\
&= \Big(1-\frac{1}{n}\Big)\sigma^2 < \sigma^2
\end{aligned}$$

In other words, the expected value of the uncorrected sample variance does not equal the population variance σ^2, unless multiplied by a normalization factor. The sample mean, on the other hand, is an unbiased estimator of the population mean μ.

Note that the unbiased sample variance S^2 can be defined as

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

We can obtain an unbiased estimator of the population variance by

$$\begin{aligned}
\mathbb{E}[S^2] &= \mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2\right] \\
&= \frac{n}{n-1}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2\right] \\
&= \frac{n}{n-1}\left(1-\frac{1}{n}\right)\sigma^2 = \sigma^2
\end{aligned}$$

9.2 Maximum Likelihood Estimation

So far, we have discussed estimating the mean and variance of a distribution. Our methods have been somewhat ad-hoc. More specifically, it is not clear how we can estimate other parameters.

We now would like to talk about a systematic way of parameter estimation. Specifically, we would like to introduce an estimation method, called maximum likelihood estimation (MLE). To give you the idea behind MLE let us look at an example.

9.2.1 Asymptotic Properties of MLEs

Let $X_1, \ldots, X_n$ be a random sample from a distribution with a parameter θ Let $\hat{\theta}_{ML}$ denote the MLE of θ. Then, under some mild regularity conditions,

- $\hat{\theta}_{ML}$ is asymptotically consistent, *i.e.,*

$$\lim_{n\to\infty} P(|\hat{\theta}_{ML} - \theta| > \epsilon) = 0$$

- $\hat{\theta}_{ML}$ is asymptotically unbiased, *i.e.,*

$$\lim_{n\to\infty} E[\hat{\theta}_{ML} - \theta] > \epsilon = \theta$$

- As n becomes large, $\hat{\theta}_{ML}$ is approximately a normal random variable. More precisely, the random variable

$$\frac{\hat{\theta}_{ML} - \theta}{\sqrt{\text{Var}(\hat{\theta}_{ML})}}$$

converges to $\mathcal{N}(0, 1)$.

9.3 Interval Estimation (Confidence Intervals)

Let $X_1, \ldots, X_n$ be a random sample from a distribution with a parameter θ that is to be estimated. Suppose that we have observed $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$. So far, we have discussed point estimation for θ. The point estimate $\hat{\theta}$ alone does not give much information about θ. In particular, without additional information, we do not know how close $\hat{\theta}$ is to the real θ. Here, we will introduce the concept of interval estimation. In this approach, instead of giving just one value $\hat{\theta}$ as the estimate for θ, we will produce an interval that is likely to include the true value of θ. Thus, instead of saying

$$\hat{\theta} = 34.25,$$

we might report the interval

$$[\hat{\theta}_l, \hat{\theta}_h] = [30, 40],$$

which we wish to include the real value of θ. We will measure two estimates for θ, a high estimate and a low estimate. There are two important concepts: First, the *length* of the reported interval, $\hat{\theta}_l = -\hat{\theta}_h$. The length of the interval shows the precision with which we can estimate θ. The smaller the interval, the higher the precision with which we can estimate θ. The second important factor is the *confidence level* that shows how confident we are about the interval. The confidence level is the probability that the interval that we construct includes the real value of θ. Therefore, higher confidence level is desirable.

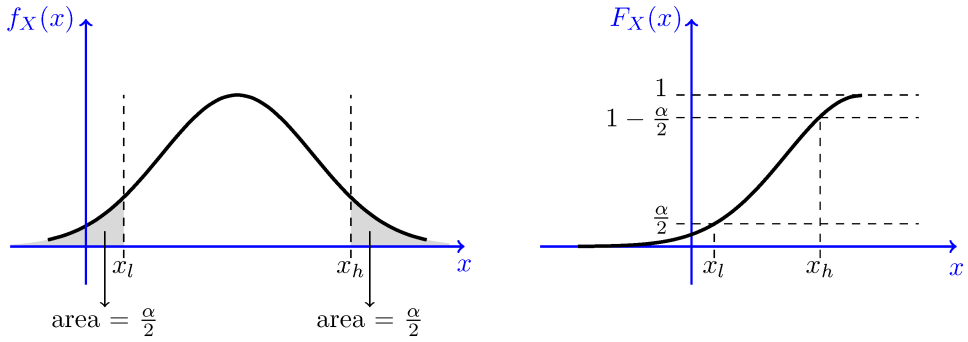


Figure 9.1: $(1-\alpha)$ interval for X.

9.3.1 The General Framework of Interval Estimation

Let $X_1, \ldots, X_n$ be a random sample from a distribution with a parameter θ that is to be estimated. Our goal is to find two estimator for θ:

- The low estimator, $\hat{\theta}_l = \hat{\theta}_l(X_1, \ldots, X_n)$
- The high estimator, $\hat{\theta}_h = \hat{\theta}_h(X_1, \ldots, X_n)$

The interval estimator is given by the interval $[\hat{\theta}_l, \hat{\theta}_h]$. The estimators are chosen such that the probability of the interval including θ is larger than $1-\alpha$. Here, $1-\alpha$ is said to be *confidence level*. We would like α to be small so that the confidence level can be higher. Common values for α are 0.1, 0.05, or 0.01 which correspond to confidence level $90\%, 95\%$, and 99%, respectively. Thus, when we are asked to find a 95% confidence interval for a parameter θ, we need to find the estimators such that

$$P\left(\hat{\theta}_l < \theta \text{ and } \hat{\theta}_h > \theta\right) \geq 0.95.$$

9.3.2 Finding Interval Estimators

Let's review a simple things first. Let X be a continuous random variable with CDF $F_X(x) = P(X \leq x)$. Suppose that we are interested in finding two values x_h and x_l such that

$$P(x_l \leq X \leq x_h) = 1 - \alpha.$$

One way to do this is to choose x_l and x_h such that

$$P(X \leq x_l) = \frac{\alpha}{2}, \text{ and } P(X \geq x_h) = \frac{\alpha}{2}.$$

Equivalently,

$$F_X(x_l) = \frac{\alpha}{2}, \text{ and } F_X(x_h) = P(X \leq x_h) = 1 - \frac{\alpha}{2}.$$

We can rewrite these equations by using the inverse function F_X^{-1} as

$$x_l = F_X^{-1}\left(\frac{\alpha}{2}\right), \text{ and } x_h = F_X^{-1}(x_h)\left(1 - \frac{\alpha}{2}\right).$$

We call the interval $[x_l, x_h]$ a $(1-\alpha)$ interval for X.

Example: Let $Z \sim \mathcal{N}(0,1)$, find x_l and x_h such that

$$P(x_l \leq Z \leq x_h) = 0.95$$

Here, $\alpha = 0.05$ and the CDF of Z is Φ (*i.e.,* standard normal distribution). Thus, we can choose

- $x_l = \Phi^{-1}(0.025) = -1.96$
- $x_h = \Phi^{-1}(1 - 0.025) = 1.96$

Thus, we have

$$P(-1.96 \leq Z \leq 1.96) = 0.95.$$

9.3.3 Confidence Intervals for Normal Samples

We assumed n to be large so that we could use the CLT. An interesting aspect of the confidence intervals that we obtained was that they often did not depend on the details of the distribution from which we obtained the random sample. That is, the confidence intervals only depended on statistics such as $\overline{X}$ and S^2. What if n is not large enough? In this case, we cannot use the CLT, so we need to use the probability distribution from which the random sample is obtained. A very important case is when we have a sample $X_1, \ldots, X_n$ from a normal distribution. Here, we would like to discuss how to find interval estimators for the mean and the variance of a normal distribution. We first need to introduce two distributions that are related to the normal distribution.

Degree of Freedom :Degrees of freedom (df) is a concept used in statistical analyses to describe the number of independent values or quantities which can vary in an analysis without breaking any constraints. Degrees of freedom can be defined as the number of independent pieces of information available to estimate another piece of information. More formally, it's the number of values in a calculation that are free to vary.

- Single Sample Mean: When estimating the mean of a sample, degrees of freedom refer to the number of observations that can vary independently. If you have n observations and you know the sample mean, $n-1$ of those observations can vary freely, while the last one is fixed by the constraint that the sum of deviations from the mean must be zero. For instance, a sample of 10 observations, the degrees of freedom when calculating the sample variance is $n - 1 = 9$. This is because once you know the first 9 values and the sample mean, the 10th value is determined.

Chi-Squared Distribution: Recall that a random variable of a gamma distribution with parameters $\alpha > 0$ and $\lambda > 0$ is given by

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

We know that if $Z_1, \ldots, Z_n$ are **independent standard normal** random variables, then the random variable

$$X = Z_1 + \cdots + Z_n$$

is also normal. More specifically, $X \sim \mathcal{N}(0, n)$. Now if we define a random variable Y as

$$Y = Z_1^2 + \cdots + Z_n^2,$$

then Y is said to have a **chi-squared** distribution with n **degrees of freedom** shown by

$$Y \sim \chi^2(x).$$

Note that the shape of the distribution depends on n. It can be shown that the random variable Y has, in fact, a gamma distribution with parameters $\alpha = \frac{n}{2}$ and $\lambda = \frac{1}{2}$,

$$Y \sim \text{Gamma}\left(\frac{n}{2}, \frac{1}{2}\right).$$

The chi-squared distribution has some properties:

- Mean: n
- Variance $2n$
- Shape: For small degrees of freedom, the chi-squared distribution is skewed to the right (long tail at the right). As the degrees of freedom increase, the distribution becomes more symmetric and approaches a normal distribution.

t-Distribution (Student's t-distribution): The Student's t-distribution, commonly referred to as the t-distribution, is a continuous probability distribution that is particularly useful in the context of small sample sizes. It is often used in hypothesis testing and constructing confidence intervals for population means when the population standard deviation is unknown.

The t-distribution arises when estimating the mean of a normally distributed population in situations where the sample size is small and the population standard deviation is unknown. Mathematically, if $X_1, X_2, \ldots, X_n$ are independent, identically distributed random variables from a normal distribution with mean μ and standard deviation σ, and $\overline{X}$ is the sample mean and S is the sample standard deviation, the *t-statistic* is given by:

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

This t-statistic follows a t-distribution with $n - 1$ degrees of freedom, where n is the sample size.

9.3.4 Confidence Intervals for the Mean of Normal Random Variables

We assume that $X_1, \ldots, X_n$ are random samples from a normal distribution $\mathcal{N}(\mu, \sigma^2)$ and our goal is to find an interval estimator for μ We no longer require n to be large. Thus, n could be any positive integer. There are two possible scenarios depending on whether σ^2 is known or not. If the value of σ^2 is known, we can easily find a confidence interval for μ. This can be done by using exactly the same method that we used to estimate μ for a general distribution for the case of large n. More specifically, we know that the random variable

$$Q = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

has $\mathcal{N}(0,1)$ distribution. In particular, Q is a function of the X_i's and μ, and its distribution does not depend on μ. Thus, Q is a pivotal quantity, and we conclude that $\left[\overline{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$ is $(1-\alpha)100\%$ confidence interval for μ.

The second case is where we do not know the variance σ^2. In this case, we can leverage the t-statistic:

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

with $n-1$ degrees of freedom.

Example:

- Using the t-Distribution for Confidence Intervals: When we don't know the exact standard deviation of a population and our sample size is small (usually less than 30), we use the t-distribution to calculate our confidence interval. Here's how you can do it step-by-step:
  - Collect Your Sample Data: Suppose you measured the heights of 20 people in your city.
  - Calculate the Sample Mean ($\overline{x}$): Add up all the heights and divide by the number of people. Let's say the average height is 170 cm.
  - Calculate the Sample Standard Deviation (s): This tells you how spread out the heights are. Let's say it's 10 cm.
  - Determine the Sample Size (n): In this case, $n = 20$.
  - Choose Your Confidence Level: Common choices are 90%, 95%, and 99%. Let's choose 95%, meaning we're 95% sure the true average height falls within our range.
  - Find the t-Value: This value comes from a t-distribution table and depends on your chosen confidence level and the degrees of freedom (which is the sample size minus one). For 20 people, degrees of freedom = 20 - 1 = 19. The t-value for 19 degrees of freedom at 95% confidence is about 2.093.
  - Calculate the Margin of Error (ME):

    $$ME = t \times \frac{s}{\sqrt{n}}$$

    Plugging in our numbers:

    $$ME = 2.093 \times \frac{10}{\sqrt{20}} \approx 4.68$$

  - Construct the Confidence Interval: Add and subtract the margin of error from the sample mean:

    Confidence Interval $= (\bar{x} - ME, \bar{x} + ME) = (170 - 4.68, 170 + 4.68) = (165.32, 174.68)$

    So, you can say with 95% confidence that the average height of people in your city is between 165.32 cm and 174.68 cm.
- Using the Chi-Squared Distribution for Confidence Intervals: The chi-squared distribution is used to find confidence intervals for variances (how spread out data is). Imagine you want to know how much people's heights vary in your city.

- Collect Your Sample Data: Again, measure the heights of 20 people.
- Calculate the Sample Variance (s^2): This is the square of the sample standard deviation. If the standard deviation is 10 cm, the variance is $10^2 = 100$ cm^2.
- Determine the Sample Size (n): Still $n = 20$.
- Choose Your Confidence Level: Let's stick with 95
- Find the Critical Chi-Squared Values: These values come from a chi-squared distribution table. You need two values: one for the lower end and one for the upper end of the confidence interval.
  - For 19 degrees of freedom (since $n - 1 = 19$):
  - The value for the lower end (0.025 in each tail for 95%) is about 32.852.
  - The value for the upper end (0.975) is about 8.907.
- Construct the Confidence Interval for the Variance:

$$\text{Lower Limit} = \frac{(n-1)s^2}{\chi^2_{\text{upper}}} = \frac{(19 \times 100)}{32.852} \approx 57.79$$

$$\text{Upper Limit} = \frac{(n-1)s^2}{\chi^2_{\text{lower}}} = \frac{(19 \times 100)}{8.907} \approx 213.33$$

So, you can say with 95% confidence that the variance in heights is between 57.79 cm^2 and 213.33 cm^2.

Chapter 10

Hypothesis Testing

A *hypothesis* is a statement that requires testing by observation to determine whether it is true or false. A few examples:

- The coin is unbiased.
- Students entering the graduate program have GPA 3.

As you can see from these examples, a hypothesis is something we can test based on the data. Therefore, being correct or wrong depends on the statistics we have and the cutoff threshold. **Accepting or rejecting a hypothesis does not mean that the statement is correct or wrong, since the truth is unknown. If we accept a hypothesis, we have made a better decision solely based on the statistical evidence**. It is possible that tomorrow when you have collected more data we may reject a previously accepted hypothesis.

The procedure for testing whether a hypothesis should be accepted or rejected is known as *hypothesis testing*. In hypothesis testing, we often have two opposite hypotheses:

- The hypothesis H_0 is called the *null hypothesis* (default hypothesis)
  - Assumed to be true
- The hypothesis H_1 is called the *alternative hypothesis* (contradictory hypothesis).

Example

Suggest a null hypothesis and an alternative hypothesis regarding whether a coin is unbiased.
Solution: Let θ be the probability of getting a head.

- $H_0 : \theta = 0.5$, and $H_1 : \theta > 0.5$. This is a **one-sided** alternative.
- $H_0 : \theta = 0.5$, and $H_1 : \theta < 0.5$. This is **another one-sided** alternative.
- $H_0 : \theta = 0.5$, and $H_1 : \theta \neq 0.5$. This is a **two-sided** alternative.

10.1 Critical-value test

In hypothesis testing, there are two major approaches:

1. **critical-value test**
2. **p-value test**.

Suppose that we have a 4-sided die and our goal is to test whether the die is unbiased. To do so, we define the null and the alternative hypotheses as

- $H_0 : \theta = 0.25$, which is our default belief.
- $H_1 : \theta > 0.25$, which is a one-sided alternative.

We must obtain data prior to conducting any hypothesis testing. Let's assume that we have thrown the die $N = 1000$ times. We find that "3" appears 290 times (we could just as well have chosen 1, 2, or 4). We let $X_1, \ldots, X_1000$ be the $N = 1000$ binary random variables representing whether we have obtained a "3" or not. If the true probability is $\theta = 0.25$, then we will have $P[X_n = 3] = \theta = 0.25$ and $P[X_n \neq 3] = 1 - \theta = 0.75$. We know that we cannot access the true probability, so we can only construct an estimator of the probability:

$$\hat{\Theta} = \frac{1}{N} \sum_{n=1}^{N} X_n.$$

In this example, we can show that $\hat{\Theta} = 290/1000 = 0.29$.

To make our problem slightly easier, we pretend that we know the variance $Var[X_n]$. In practice, we certainly do not know $Var[X_n]$, and so we need to estimate the variance. If we knew the variance, it should be $Var[X_n] = \theta(1 - \theta) = 0.25(1 - 0.25) = 0.1875$, because X_n is a Bernoulli random variable with a mean θ.

The question asked by hypothesis testing is: How far is "$\hat{\Theta} = 0.29$" from "$\theta = 0.25''$?

- If the statistic generated by our data, $\hat{\Theta} = 0.29$, is "far" from the hypothesized $\theta = 0.25$, then we need to reject H_0 because H_0 says that $\theta = 0.25$.
- However, if there is no strong evidence that $\theta > 0.25$, we will need to assume that H_0 may possibly be true. So the key question is what is meant by "far".

For many problems like this one, it is possible to analyze the PDF of $\hat{\Theta}$. Since $\hat{\Theta}$ is the sample average of a sequence of Bernoulli random variables, it follows that $\hat{\Theta}$ is a binomial (with a scaling constant $1/N$). If N is large enough, e.g., $N \geq 30$, the Central Limit Theorem tells us that $\hat{\Theta}$ is also very close to a Gaussian. Therefore, we can more or less claim that

$$\hat{\Theta} \sim \mathcal{N}\left(\theta, \frac{\sigma^2}{N}\right).$$

With a simple translation and scaling, we can normalize $\hat{\Theta}$ to obtain $\hat{Z}$:

$$\hat{Z} = \frac{\hat{\Theta} - \theta}{\sigma/\sqrt{N}} \sim \mathcal{N}(0,1)\,.$$

One essential element of hypothesis testing is the *cutoff threshold*, which is defined through the *critical level* α. It is the area under the curve of the PDF of $\hat{Z}$. Typically, α is chosen to be a small value, such as $\alpha = 0.05$ (corresponding to a 5% margin). The corresponding cutoff is known as the *critical value* z_α, which is defined as

$$z_\alpha\text{: cutoff location where area under the curve is } \alpha\ .$$

If $\hat{Z}$ is $\mathcal{N}(0,1)$ and if we are looking at the right-hand tail, it follows that

$$z_\alpha = \Phi^{-1}(1-\alpha).$$

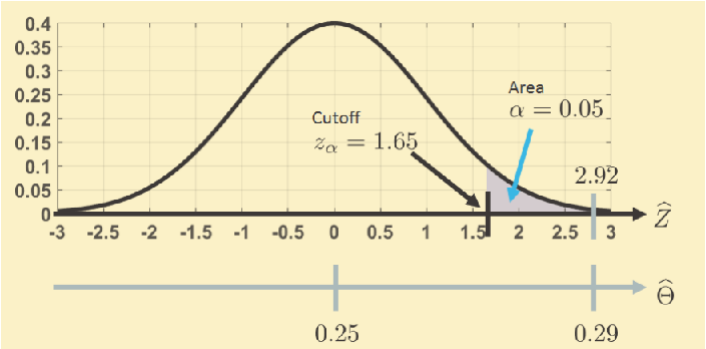


Figure 10.1: There are two axes: the $\hat{\Theta}$-axis (which is the estimator) and the $\hat{Z}$-axis (which is the normalized variable). The values corresponding to each axis are shown in the figure. For example. $\hat{\Theta} = 0.29$ is equivalent to $\hat{Z} = 2.92$, and $\hat{\Theta} = 0.25$ is equivalent to $\hat{Z} = 0$, etc. Therefore, when we ask how far "$\hat{\Theta} = 0.29$" is from "$\theta = 0.25$", we can map this question from the $\hat{\Theta}$-axis to the $\hat{Z}$-axis, and ask the relative position of $\hat{Z}$ from the origin.

```
# Etimate the Z_hat value
import numpy as np
Theta_hat = 0.29 # Your estimate
theta     = 0.25 # Your hypothesis
N         = 1000 # Number of samples
sigma     = np.sqrt(theta*(1-theta)) # Known standard deviation

Z_hat = (Theta_hat - theta)/(sigma / np.sqrt(N))
print(Z_hat)
```

To determine the critical value z_α,

```
# Compute the critical value
import scipy.stats as stats
alpha = 0.05
z_alpha = stats.norm.ppf(1-alpha, 0, 1)
```

- PPF stands for *percent point function*, which is an inverse of CDF.

Do we have enough evidence to reject H_0 in this example? Yes!

- The estimated value $\hat{\Theta} = 0.29$ is equivalent to $\hat{Z} = 2.92$, which is much too far from the cutoff $z_\alpha = 1.65$.
- In other words, we conclude that at a 5% critical level we have strong evidence to believe that the die is biased. Therefore, we need to reject H_0.

The decision based on comparing the critical value is known as the *critical-value test*.

1. Set a critical value z_α. Then, compute
$$\hat{Z} = \frac{\hat{\Theta} - \theta}{\sigma/\sqrt{N}}$$
2. If $\hat{Z} \geq z_\alpha$, then reject H_0.
3. If $\hat{Z} < z_\alpha$, then keep H_0.

The critical-value test belongs to a larger family of testing procedures based on decision theory. To give you a preview of the general theory of hypothesis testing, we define a *decision rule*, a function that maps a realization of the estimator to a binary decision space. In our problem the estimator is $\hat{Z}$ (or equivalently $\hat{\Theta}$). We denote its realization by $\hat{z}$. The binary decision space is $\{H_0, H_1\}$, corresponding to whether we want to claim H_0 or H_1. Claiming H_0 is equivalent to keeping H_0, and claiming H_1 is equivalent to rejecting H_0. For the critical-value test, the decision rule $\delta(\cdot) : \mathbb{R} \to \{0, 1\}$ is given by the equation (for testing a right-hand tail):

$$\delta(\hat{z}) = \begin{cases} 1 & \text{if } \hat{z} \geq z_\alpha \text{ claim } H_1. \\ 0 & \text{if } \hat{z} < z_\alpha \text{ claim } H_0. \end{cases}$$

10.2 p-value test

An alternative to the critical-value test is the *p-value test*. Instead of looking at the cutoff value z, **we inspect the probability of obtaining our observation if H_0 is true**. To understand how the p-value test works, we consider another toy problem.

Suppose that we have two hypotheses about flipping a coin:

- $H_0 : \theta = 0.9$, which is our default belief.
- $H_1 : \theta < 0.9$, which is a one-sided alternative.

It was found that with $N = 150$ coin flips, the coin landed on heads 128 times. Thus the estimator is $\hat{\Theta} = \frac{128}{150} = 0.853$. Then, by following our previous procedures, we have that

$$\hat{Z} = \frac{\hat{\Theta} - \theta}{\sigma/\sqrt{N}} = \frac{0.853 - 0.9}{\sqrt{\frac{0.9(1-0.9)}{150}}} = -1.92.$$

At this point we can follow the previous subsection by computing the critical value z and make the decision. However, let's take a different route. We want to know what is the probability

under the curve if we integrate the PDF of $\hat{Z}$ from $-\infty$ to -1.92. This is easy. Since $\hat{z}$ is $\mathcal{N}(0,1)$, it follows from the CDF of a Gaussian that

$$\underbrace{\mathbb{P}[\hat{Z} \leq -1.92]}_{p\text{-value}} = 0.0274.$$

The value 0.0274 is the pink area under the curve, which is the PDF of $\hat{Z}$. Since the area under the curve is less than the critical level α (say 5%), we reject the null hypothesis.

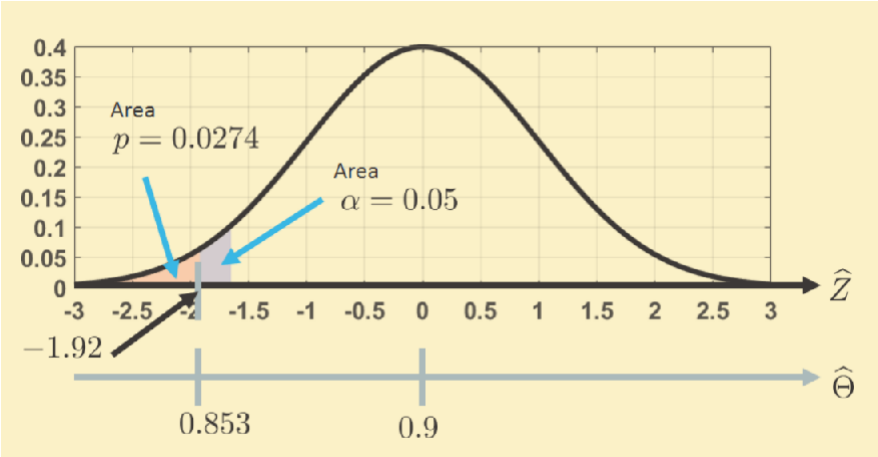


Figure 10.2: The p-value test asks us to look at the probability of $\hat{Z} \leq \hat{z}$. If this probability (the p-value) is less than the critical level α, we have significant evidence to reject the null hypothesis.

```
# Compute the p-value
import scipy.stats as stats
p = stats.norm.cdf(-1.92,0,1)
```

In this example, the probability $P[\hat{Z} \leq -1.92]$ is known as the p-value. It is the probability of $\hat{Z} \leq \hat{z}$, under the distribution mandated by the null hypothesis, where z is the (normalized) estimated value based on data. Using our example, $zis - 1.92$. By "distribution mandated by the null hypothesis" we mean that the PDF of $\hat{z}$ is the PDF that the null hypothesis wants. In the above example the PDF is $Gaussian(0,1)$, corresponding to $Gaussian(\theta, \sigma/\sqrt{N})$ for $\hat{\Theta}$.

The decision rule based on the p-value is

$$\delta(\hat{z}) = \begin{cases} 1 & \text{if } \mathbb{P}[\hat{Z} \leq \hat{z}] < \alpha \text{ claim } H_1. \\ 0 & \text{if } \mathbb{P}[\hat{Z} \leq \hat{z}] \geq \alpha \text{ claim } H_0. \end{cases}$$

If the alternative hypothesis is right-handed, then the probability becomes $\mathbb{P}[\hat{Z} \geq \hat{z}]$ instead.

Example

We flip a coin for $N = 150$ times and find that 128 are heads. Consider two hypotheses

- $H_0 : \theta = 0.9$, which is our default belief.
- $H_1 : \theta \neq 0.9$, which is a two-sided alternative.

For a critical level of $\alpha = 0.05$, shall we keep or reject H_0?

We know that $\hat{\Theta} = 128/150 = 0.853$. The normalized statistic is

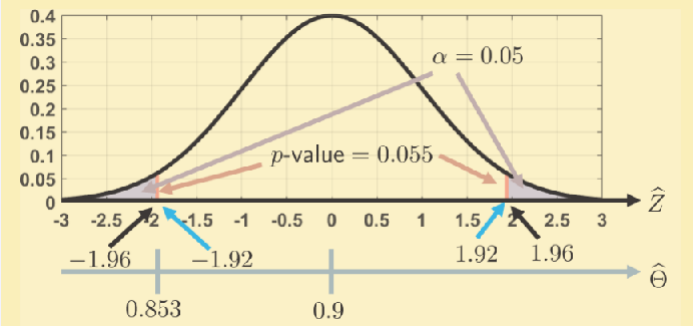
$$\hat{Z} = \frac{\hat{\Theta} - \theta}{\sigma/\sqrt{N}} = \frac{0.853 - 0.9}{\sqrt{\frac{0.9(1-0.9)}{150}}} = -1.92.$$

To compute the p-value, we observe that the two-sided test means that we consider the two tails. Thus, we have

$$\begin{aligned} p - \text{value} &= \mathbb{P}[|\hat{Z}| \geq 1.92] \\ &= 2 \times \mathbb{P}[\hat{Z} \geq 1.92] \\ &= 2 \times 0.0274 = 0.055. \end{aligned}$$

For a critical level of $\alpha = 0.05$, the p-value is larger. This means that the probability of obtaining $|Z| > 1.92$ is not extreme enough. Therefore, we do not have sufficient evidence to reject the null hypothesis.
If we take the critical-value test, we will reach the same conclusion. The critical value for $\alpha = 0.05$ is determined by taking the inverse CDF at $1 - 0.025$, giving

$$z_\alpha = \Phi^{-1}(1 - \frac{\alpha}{2}) = 1.96.$$



10.2.1 p-value hacking

p-value hacking, also known as *data dredging*, data fishing, data snooping or data butchery, is **an exploitation of data analysis in order to discover patterns which would be presented as statistically significant, when in reality, there is no underlying effect**. In other words, p-hacking is **running statistical tests on a set of data until some statistically significant results arise**. That can be done in a few different ways, for example: by stopping the collection of data once you get a $p < 0.05$ (reject null hypothesis to claim their hypothesis), analyzing many outcomes, but only reporting those with $p < 0.05$, using covariates, excluding participants, etc.

10.3 Z-test and T-test

The critical-value test and the p-value tests are generic tools for hypothesis testing. In this subsection we introduce the Z-test and the T-test. It is important to understand that the Z-test and the T-test refer to the distributional assumptions we make about the variance. They define the distribution we use to conduct the test but not the tools. In fact, both the Z-test and the

T-test can be implemented using the critical-value test or the p-value test.

The difference between the Gaussian distribution and the T distribution is mainly attributable to the knowledge about **the population variance**:

- If the variance is known, the distribution of the estimator (which in our case is the sample average) is **Gaussian**.
- If the variance is estimated from the sample, the distribution of the estimator will follow a Student's **t-distribution**.

10.4 Neyman-Pearson Test

10.4.1 Null and alternative distributions

When we discussed hypothesis testing in the previous section, we focused exclusively on the null hypothesis H_0. Regardless of whether we are studying the Z-test or the T-test, using the critical value or the p-value, all the distributions are associated with the distribution under H_0.

What do we mean by "distribution under H_0"? Using $\hat{\Theta}$ as an example, the PDF of $\hat{\Theta}$ is assumed to be $\mathcal{N}(\theta, \sigma^2/N)$. This Gaussian, centered at θ, is the distribution assumed under H_0. As we decide whether to keep or reject H_0, we look at the critical value and the p-value of the test statistic under $\mathcal{N}(\theta, \sigma^2/N)$. Importantly, the analysis of hypothesis testing is not just about H_0 - it is also about the alternative hypothesis H_1, which uses a different PDF. For example, H_1 could use $\mathcal{N}(\theta', \sigma^2/N)$ for $\theta' > \theta$. Therefore, for the same testing statistic $\hat{\Theta}$, we can check how close it is to H_1.

To capture both distributions, we define

- $f_0(y) = f_Y(y|H_0)$
- $f_1(y) = f_Y(y|H_1)$

The first PDF defines the distribution when the true model is H_0. The second PDF is the distribution when the true model is H_1.

$$\delta(y) = \begin{cases} 1 & \text{if } y \in R_\alpha \text{ reject } H_1. \\ 0 & \text{if } y \notin R_\alpha \text{ keep } H_0. \end{cases}$$

Example

Consider $H_0 : \theta = 0.35$ and $H_1 : \theta > 0.35$. It was found that the sample average over 1009 samples is $\hat{\Theta} = 0.387$, with $\sigma^2 = 0.227$. The normalized test statistic is $\hat{Z} = \sqrt{N}(\hat{\Theta} - \theta)/\sigma = 2.432$. At a 5% critical level, define the decision rule based on the critical-value approach.
If $\alpha = 0.05$, it follows that $z_\alpha = \Phi^{-1}(1 - 0.05) = 1.65$, which is bigger than 1.65, so we reject H_0.

10.4.2 Type 1 and type 2 errors

- FPR: Type 1 error (declare H_1, truth is H_0), the number of FP divided by the total number of cases.
- FNR: Type 2 error (declare H_0, truth is H_1), the number of FN divided by the total number of cases.

The center is the H_1, so the positive means you declare the H_1, and the negative is the case, where you reject H_1.

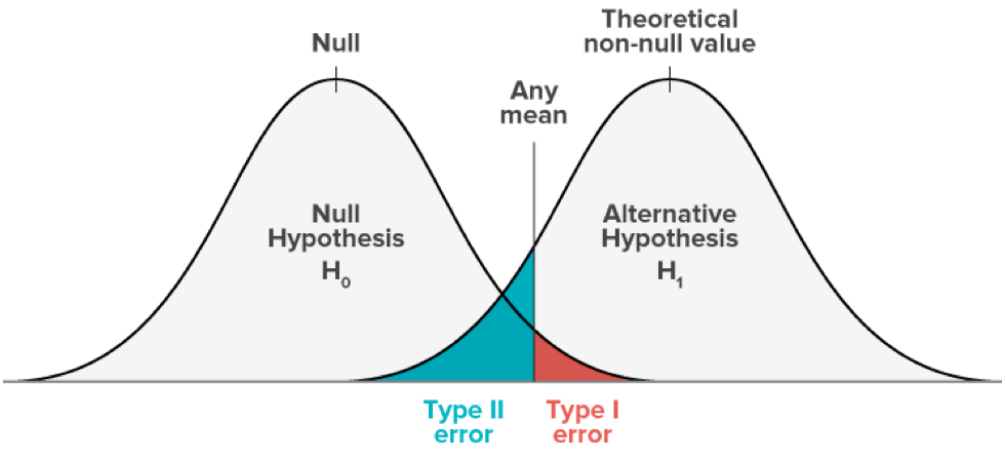


Figure 10.3: The null hypothesis says that the test result says you have no disease and the alternative hypothesis says you have disease. Then, you set a threshold at the line. Type 1 error is the case where you are diagnosed with positive, but actually you don't have the disease.

10.5 ROC and Precision-Recall Curve

Being a binary decision rule, the hypothesis testing procedure shares many similarities with a two-class classification algorithm. Given a testing statistic or a testing sample, both the hypothesis testing and a classification algorithm will report YES or NO. Therefore, any performance evaluation metric developed for hypothesis testing is equally applicable to classification and vice versa.

The topic we study in this section is the **receiver operating characteristic** (ROC) curve and the **precision-recall** (PR) curve. The ROC curve and the PR curve are arguably the most popular metrics in modern machine learning, in particular for classification, detection, and segmentation tasks in computer vision. There are many unresolved questions about these two curves and there are many debates about how to use them. Our goal is not to add another voice to the debate; rather, we would like to fill in the gap between the hypothesis testing theory (particularly the Neyman-Pearson framework) and these two sets of curves. We will establish the equivalence between the two curves and leave the open-ended debates to you.

10.5.1 Receiver Operating Characteristic (ROC)

Chapter 11

Regression

11.1 Linear Regression

Sometimes we are interested in obtaining a simple model that explains the relationship between two or more variables. For example, suppose that we are interested in studying the relationship between the income of parents and the income of their children in a certain country. In general, many factors can impact the income of a person. Nevertheless, we suspect that children from wealthier families generally tend to become wealthier when they grow up. Here, we can consider two variables:

- The family income can be defined as the average income of parents at a certain period.
- The child income can be defined as his/her average income at a certain period (e.g, age).

To examine the relationship between the two variables, we collect some data

$$(x_i, y_i), \quad \text{for } i = 1, \ldots, n,$$

where y_i is the average income of the i-th child and x_i is the average income of his/her parents. We are often interested in finding a simple model. A linear model is probably the simplest model that we can define, where we write

$$y_i = \beta_0 + \beta_1 x_i.$$

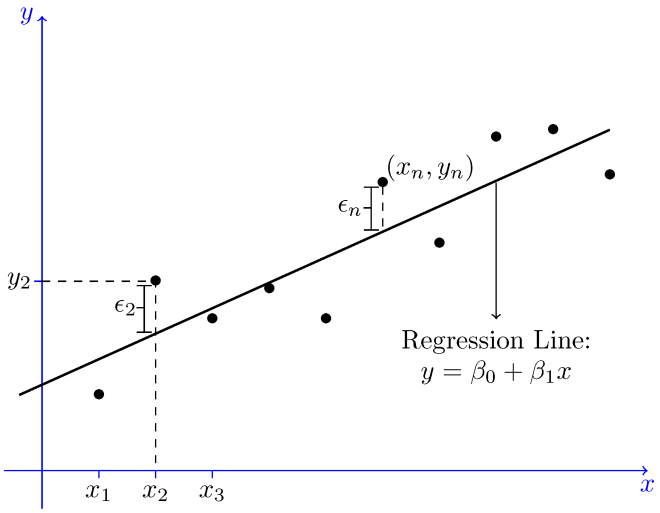
Since, there could be other factors that impact a child's future income, we can express it

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

where ϵ_i is modeled as a random variable. More specifically, if we approximate a child's future income by the linear model, then ϵ_i indicates the error in our approximation. The goal is to obtain the best values of the parameters (*regression coefficients*). Since ϵ (residual term) is a random variable, Y is also a random variable. The variable x is called the *predictor* or the *explanatory* variable, and the random variable Y is called the *response* variable.

11.1.1 A Simple Linear Regression Model

$$Y = \beta_0 + \beta_1 X + \epsilon,$$



where ϵ's as independent and zero-mean normal random variables,

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2).$$

The parameters and σ^2 are considered fixed but unknown. The assumption is that we have data points and our goal is to find the best values for the regression coefficients.

First, we take expectation from both sides to obtain

$$\begin{aligned} EY &= \beta_0 + \beta_1 EX + E[\epsilon] \\ &= \beta_0 + \beta_1 EX \end{aligned}$$

Thus,

$$\beta_0 = EY - \beta_1 EX$$

Subsequently, we look at $Cov(X, Y)$,

$$\begin{aligned} Cov(X, Y) &= Cov(X, \beta_0 + \beta_1 X + \epsilon) \\ &= \mathrm{Cov}(X, \beta_0) + \mathrm{Cov}(X, \beta_1 X) + \mathrm{Cov}(X, \epsilon) \\ &= \mathrm{Cov}(X, \beta_1 X) \\ &= \beta_1 \mathrm{Cov}(X, X) \quad \text{by Scaling property} \\ &= \beta_1 \mathrm{Var}(X) \end{aligned}$$

Note that the second step uses the linearity of covariance:

$$Cov(X + Y, Z) = Cov(X, Z) + Cov(Y, Z).$$

Also, the covariance with a constant and any random variable is zero, and the covariance of two independent random variables is zero.

Therefore, we obtain

$$\beta_1 = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)}, \quad \beta_0 = EY - \beta_1 EX.$$

In sum, if we know the above statistics, we can get the best coefficients.

Coefficient of Determination (R-Squared): Let's look at the linear mode:

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

Note that if we want to estimate Y using the observed X, then we can express it

$$\hat{Y} = \beta_0 + \beta_1 X.$$

The error in our estimate is

$$Y - \hat{Y} = \epsilon.$$

Note that the randomness in Y comes from two sources: (i) X and (ii) ϵ. More specifically, if we look at $Var(Y)$,

$$Var(Y) = \beta_1^2 Var(X) + Var(\epsilon),$$

since X and ϵ are assumed to be independent. Also, recall that $Var(aX) = a^2 Var(X)$. From the above equation we can define

$$\rho^2 = \frac{\beta_1^2 Var(X)}{Var(Y)}$$

as the portion of variance of Y that is explained by variation in X. In other words, we want to measure the impact of the variance X. The R^2 value also can be defined as follows:

$$R^2 = 1 - \frac{Var(\epsilon)}{Var(Y)}$$

Thus, it always lie between 0 and 1, since $Var(Y)$ is the total variance. Equivalently, we can leverage the fact that $\beta_1 = \frac{\text{Cov}(X,Y)}{\text{Var}(X)}$. Then, we get

$$\rho^2 = \frac{Cov(X,Y)^2}{Var(X)Var(Y)}$$

Larger values of R^2 generally suggest that our linear model is a good fit for the data.

Chapter 12

Statistical Inference: Bayesian Methods

In this chapter, we would like to discuss a different framework for inference, namely the *Bayesian* approach. In the Bayesian framework, we treat the unknown quantity, θ, as a random variable. More specifically, *we assume that we have some initial guess about the distribution of* θ. This distribution is called the *prior* distribution. After observing some data, we update the distribution of θ (based on the observed data). This step is usually done using Bayes' Rule. That is why this approach is called the Bayesian approach. The details of this approach will be clearer as you go through the chapter.

Example: Suppose that you would like to estimate the portion of voters in your town that plan to vote for Party A in an upcoming election. To do so, you take a random sample of size n from the likely voters in the town. Since you have a limited amount of time and resources, your sample is relatively small. Specifically, suppose that $n = 20$. After doing your sampling, you find out that 6 people in your sample say they will vote for Party A.

Let θ be the true portion of voters in your town who plan to vote for Party A. You might want to estimate θ as

$$\theta = \frac{6}{20} = 0.3$$

In fact, in absence of any other data, that seems to be a reasonable estimate. However, you might feel that $n = 20$ is too small. Thus, your guess is that the error in your estimation might be too high. While thinking about this problem, you remember that **the data from the previous election is available to you**. You look at that data and find out that, in the previous election, 40% of the people in your town voted for Party A. How can you use this data to possibly improve your estimate of θ? You might argue as follows: Although the portion of votes for Party A changes from one election to another, the change is not usually very drastic. Therefore, given that in the previous election 40% of the voters voted for Party A, you might want to model the portion of votes for Party A in the next election as a random variable θ with a probability density function, $f_\theta(\theta)$, that is mostly concentrated around $\theta = 0.4$. For example, you might want to choose the density such that

$$E[\theta] = 0.4.$$

Therefore, the initial belief or prior distribution is $f_\theta(\theta)$.

12.1 Maximum A Posteriori (MAP) Estimation

The posterior distribution $f_{X|Y}(x|y)$ (or $P_{X|Y}(x|y)$) contains all the knowledge about the unknown quantity X. Therefore, we can use the posterior distribution to find point or interval estimates of X. One way to obtain a point estimate is to choose the value of x that maximizes the posterior PDF or PMF. This is called MAP estimation.

To find the MAP estimate, we need to find the value of x that maximizes

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

Note that $f_Y(y)$ does not depend on the value of x. Therefore, we can use the numerator to find the MAP estimate:

$$\max_x f_{Y|X}(y|x)f_X(x).$$

This makes the problem simple, since it is difficult to compute the denominator, which typically involves computing a law of total probability.

Chapter 13

Random Process

In modern data science, many problems involve time. The stock market changes every minute; a speech signal changes every millisecond; a car changes its steering angle constantly; the examples are endless. A common theme among all these examples is randomness. We do not know whether a stock will go up or down tomorrow, although we may be able to make some predictions based on previous observations. We do not know the next word of a sentence, but we can guess based on the context. Random processes are tools that can be applied to these situations. We treat a random process as an infinitely long vector of random variables where the correlations between the individual variables define the statistical properties of the process. If we can determine these correlations, we will be able to summarize the past and predict the future.

13.1 Basic Concepts

What is a *random process*? A **random process is a function indexed by a random key**.

- Random Variable: Let X be the outcome of rolling a six-sided die. X can take values $\{1, 2, 3, 4, 5, 6\}$ with equal probability $\frac{1}{6}$.
- Random Process: Let $\{X(t) : t \in [0, \infty)\}$ be the position of a particle undergoing Brownian motion. $X(t)$ represents the position at time t, with $X(0) = 0$ and increments $X(t) - X(s)$ for $t > s$ following a normal distribution with mean 0 and variance $t - s$.

13.2 Statistical and Temporal Perspectives

Since a random process is a function indexed by a random key, it is a two-dimensional object. It is a function both of time t and of the random key ξ. That's why we use the notation $X(t, \xi)$ to denote a random process.

Temporal Perspective: Let us fix the random key at $\xi = \xi_0$. This gives us a function $X(t, \xi_0)$. Since ξ is already fixed at ξ_0, we are looking at a particular realization drawn from the sample space. This realization is expressed as a function $X(t, \xi_0)$, which is just a deterministic function that evolves over time. There is no randomness associated with it. This is analogous to a random

variable. While X itself is a random variable, by fixing the random key $\xi = \xi_0$, $X(\xi_0)$ is just a real number. For random processes, $X(t, \xi_0)$ now becomes a function.

Since $X(t, \xi_0)$ is a function that evolves over time, we view it along the horizontal axis. For example, we can study the sequence

$$X(t_1, \xi_0), X(t_2, \xi_0), \ldots, X(t_K, \xi_0),$$

where $t_1, \ldots, t_k$ are the time indices of the function. This sequence is deterministic and is just a sequence of numbers, although the numbers evolve as t changes.

Statistical Perspective: The other perspective, which could be slightly more abstract, is the statistical perspective. Let us fix the time at $t = t_0$. The random key ξ can take anystate defined in the sample space. So if the sample space contains $\{\xi_1, \ldots, \xi_N\}$, the sequence $\{X(t_0, \xi_1), X(t_0, \xi_2), \ldots, X(t_0, \xi_N), \}$ is a sequence of random variables, because the ξ's can go from one state to another state.

A good way to visualize the statistical perspective is the vertical perspective in which we write the sequence as a vertical column of random variables:

$$\begin{gathered} X(t_0, \xi_1) \\ X(t_0, \xi_2) \\ \vdots \\ X(t_0, \xi_N) \end{gathered}$$

That is, if you fix the time at $t = t_0$, you are getting a sequence of random variables. The probability of getting a particular value $X(t_0)$ depends on which random state you land on.

Why do we bother to differentiate the temporal perspective and the statistical perspective? The reason is that the operations associated with the two are different, even if sometimes they give you the same result. For example, if we take the temporal average of the random process, we get a number:

$$\overline{X}(\xi) = \frac{1}{T} \int_0^T X(t, \xi) dt.$$

We call this the "*temporal average*", because we have integrated the function over time. The resulting value will not change with time. However, $\overline{X}(\xi)$ depends on the random key you provide. If you pick a different random realization, $\overline{X}(\xi)$ will take a different value. So the **temporal average is a random variable**.

On the other hand, if we take the statistical average of the random process, we get

$$\mathbb{E}[X(t)] = \int_\Omega X(t, \xi) p(\xi) d\xi,$$

where $p(\xi)$ is the PDF of the random key ξ. We call this the *statistical average*, because we have taken the expectation over all possible random keys. The resulting object $\mathbb{E}[X(t)]$ is **deterministic** but a function of time.

No matter how you look at the temporal average or the statistical average, they are different with the following exception: that $\overline{X}(\xi) = \mathbb{E}[X(t)] = a$. This happens only for some special random processes, known as *ergodic* processes that allow us to approximate the statistical average using the temporal average, wish some guarantees derived from the law of large numbers.

13.3 Mean and Correlation Functions

13.3.1 Mean Function

The mean function $\mu_X(t)$ of a random process $X(t)$ is

$$\mu_X(t) = \mathbb{E}[X(t)]$$

Let's consider the "expectation" of $X(t)$. Recall that a random process is actually $X(t, \xi)$ where ξ is the random key. Therefore, the expectation is taken with respect to ξ, or formally,

$$\mu_X(t) = \mathbb{E}[X(t)] = \int_\Omega X(t, \xi) p(\xi) d\xi,$$

where $p(\xi)$ is the PDF of the random key and Ω is the sample space of ξ.

13.3.2 Autocorrelation Function

In random processes, the notions of variance and covariance are trickier than for random variables. Let us first define the concept of an *autocorrelation function* of a random process $X(t)$, which is given by

$$R_X(t_1, t_2) = \mathbb{E}[X(t_1)X(t_2)].$$

How do we understand the meaning of $\mathbb{E}[X(t_1)X(t_2)]$? $\mathbb{E}[X(t_1)X(t_2)]$ is analogous to the correlation $\mathbb{E}[XY]$ between two random variables. The $\mathbb{E}[XY]$ could be regarded as the inner product of two vectors, and so it is a measure of the closeness between X and Y. Now, if we substitute X and Y with $X(t_1)$ and $X(t_2)$ respectively, then we are effectively asking about the closeness between $X(t_1)$ and $X(t_2)$. So, in a nutshell, **the autocorrelation function tells us the correlation between the function at two different time stamps**.

What do we mean by the correlation between two timestamps? Recall that $X(t_1)$ and $X(t_2)$ are two random variables. Consider the following example:

Example: Let $X(t) = A\cos(2\pi t)$, where $A \sim Unif[0, 1]$. Find $\mathbb{E}[X(0)X(0.5)]$.

If $X(t) = A\cos(2\pi t)$, then

- $X(0) = A\cos(0) = A$
- $X(0.5) = A\cos(0.5) = -A$

Then, we get

$$\begin{aligned}\mathbb{E}[X(0)X(0.5)] &= -\mathbb{E}[A \cdot A] \\ &= -\mathbb{E}[A^2] = -\frac{1}{3}\end{aligned}$$

If we consider $X(0)$ and $X(0.5)$, each of them is a random variable, and thus we can ask about their PDFs. It is clear that $X(0)$ has a PDF that is a uniform distribution from 0 to 1 by its definition, whereas $X(0.5)$ follows a uniform distribution from -1 to 0. Formally,

$$f_{X(0)} = \begin{cases} 1 & 0 \le x \le 1, \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad f_{X(0.5)} = \begin{cases} 1 & -1 \le x \le 0, \\ 0 & \text{otherwise} \end{cases}$$

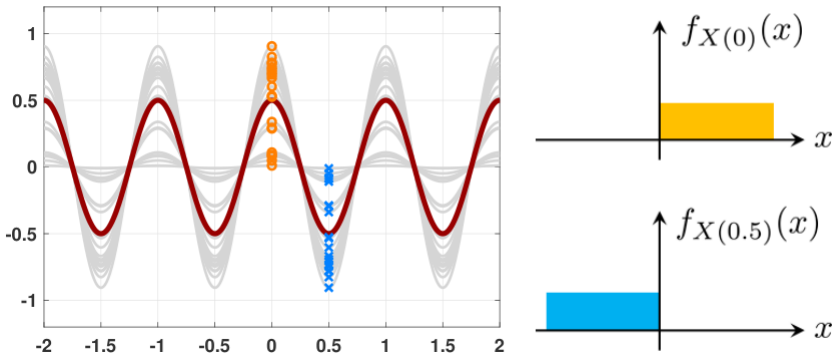


Figure 13.1: The autocorrelation between $X(0)$ and $X(0.5)$ should be regarded as the correlation between two random variables. Each random variable has its own PDF.

13.3.3 Independent Processes

How do we establish independence for two random processes? We know that for two random variables to be independent, the joint PDF can be written as a product of two PDFs:

$$f_{X,Y}(x,y) = f_X(x)f_Y(y).$$

If we extrapolate this idea to random processes, a natural formulation would be

$$f_{X(t),Y(t)}(x,y) = f_{X(t)}(x)f_{Y(t)}(y).$$

However, this definition has a problem since $X(t)$ and $Y(t)$ are functions. It is not enough to just look at one time index, say $t = t_0$. The way to think about this situation is to consider a pair of random vectors $\mathbf{X}$ and $\mathbf{Y}$. When you say they are independent, you require $f_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y}) = f_{\mathbf{X}}(\mathbf{x})f_{\mathbf{Y}}(\mathbf{y})$. The PDF $f_{\mathbf{X}}(\mathbf{x})$ itself is a joint distribution, *i.e.*, $f_{\mathbf{X}}(\mathbf{x}) = f_{X_1,\ldots,X_N}(x_1,\ldots,x_N)$. Therefore, for random processes, we need something similar.

Definition 3 Two random processes $X(t)$ and $Y(t)$ are independent if for any $t_1,\ldots,t_N$,

$$\begin{aligned} f_{X(t_1),\ldots,X(t_N),Y(t_1),\ldots,Y(t_N)}(x_1,\ldots,x_N,y_1\ldots,y_N) \\ = f_{X(t_1),\ldots,X(t_N)}(x_1,\ldots,x_N)f_{Y(t_1),\ldots,Y(t_N)}(y_1,\ldots,y_N) \end{aligned}$$

Independence means that the behavior of one process will not influence the behavior of the other process. We define *uncorrelated* as follows: Two random processes, $X(t)$ and $Y(t)$ are uncorrelated if

$$\mathbb{E}[X(t_1)Y(t_2)] = \mathbb{E}[X(t_1)]\mathbb{E}[Y(t_2)].$$

Independence implies uncorrelation, but if two random processes are uncorrelated, they are not necessarily independent.

13.4 Wide-Sense Stationary Processes

Random processes with Toeplitz structure are known as *wide-sense stationary* (WSS) processes. WSS processes belong to a very small subset in the entire universe of random processes, but they are practically the most useful ones. Before we discuss how to use them, we first present a formal definition of a WSS process.

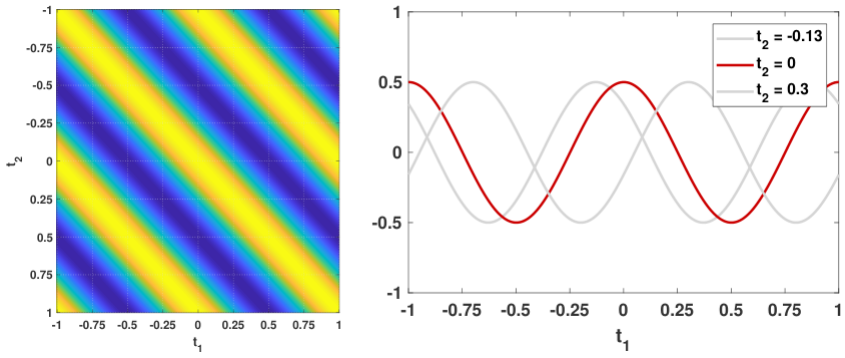


Figure 13.2: Cross sections of the autocorrelation function $R_X(t_1, t_2) = \frac{1}{2}\cos(\omega(t_1 - t_2))$.

Definition 4 *WSS Process* A random process $X(t)$ is wide-sense stationary if

1. $\mu_X(t) = constant$, for all t
2. $R_X(t_1, t_2) = R_X(t_1 - t_2)$, for all t_1, t_2.

The first criterion states that **the mean function does not change with time**. The second criterion is that the autocorrelation function only depends on the difference $t_1 - t_2$ and not on the absolute starting point. For example, $R_X(0.1, 1.1)$ needs to be the same as $R_X(6.3, 7.3)$, since the intervals are both 1.

How can these two criteria be mapped to the Toeplitz structure? The figure above represents the autocorrelation function $R_X(t_1, t_2)$, which is a 2D function. We take three cross sections corresponding to $t_2 = -0.13, t_2 = 0$, and $t_2 = 0.3$. As you can see from the figure, each $R_X(t_1, t_2)$ is a shifted version of another. To obtain any value $R_X(t_1, t_2)$ on the function, there is no need to probe to the 2D map; you only need to probe to the red curve and locate the position marked as $t_1 - t_2$, and you will be able to obtain the value $R_X(t_1, t_2)$. Since a WSS is completely characterized by the difference $t_1 - t_2$, there is no need to keep track of the absolute indices t_1 and t_2. We can rewrite the autocorrelation function as

$$R_X(\tau) = \mathbb{E}[X(t + \tau)X(t)].$$

Part II

Advanced Topics

Bibliography

[1] H. Pishro-Nik. Introduction to Probability, Statistics, and Random Processes. https://www.probabilitycourse.com/, 2024. Online; accessed 29 June 2024.

[2] Chan Stanley H. *Introduction to Probability for Data Science*. Michigan Publishing, Michigan, 2021.