Probability & Random Processes

Sara Pohland

Created: January 29, 2021

Last Modified: December 14, 2023

Contents

Part I

Introduction to Probability

Chapter 1

Experiments, Models, & Probability

1.1 Probability Spaces

A probability space $(\Omega, \mathcal{F}, P)$ is a mathematical construct that allows us to model an experiment. The **sample space** Ω is the set of all possible outcomes/observations of the experiment. The outcomes must be mutually exclusive and collectively exhaustive. The **event space** $\mathcal{F}$ denotes the set of all possible events, where each event is a subset of the sample space containing zero or more outcomes. The **probability law** $P : \mathcal{F} \to [0, 1]$ assigns a probability between zero and one to each event in the event space.

Example: Consider tossing a fair coin twice. Each toss, we can get either head (H) or tail (T). Suppose we are interested in how many times we get a head.

$$\Omega = \{HH, HT, TH, TT\}$$

$$\mathcal{F} = \{n \text{ heads}, \ n \in \mathbb{Z}\}$$

$$P(n \text{ heads}) = \begin{cases} 1/4 & \text{if } n = 0 \\ 1/2 & \text{if } n = 1 \\ 1/4 & \text{if } n = 2 \\ 0 & \text{otherwise} \end{cases}$$

A probability law must satisfy the following **probability axioms**:

1. Non-negativity – $P(A) \geq 0, \ \forall A \in \mathcal{F}$
2. Normalization – $P(\Omega) = 1$
3. Additivity – $P\big(\cup_i A_i\big) = \sum_i P(A_i), \ \forall$ disjoint $A_i \in \mathcal{F}$

Probability laws also satisfy several other properties that can be derived from the probability axioms:

1. $P(\emptyset) = 0$
2. $P(A^C) = 1 - P(A),\ \forall A \in \mathcal{F}$
3. $A \subset B \implies P(A) \leq P(B),\ \forall A, B \in \mathcal{F}$
4. $P(A \cup B) = P(A) + P(B) - P(A \cap B),\ \forall A, B \in \mathcal{F}$
5. $P(A \cup B \cup C) = P(A) + P(A^C \cap B) + P(A^C \cap B^C \cap C),\ \forall A, B, C \in \mathcal{F}$
6. $P\Big(\bigcup_{i=1}^{n} A_i\Big) \leq \sum_{i=1}^{n} P(A_i),\ \forall A_i \in \mathcal{F}$

1.2 Conditional Probability & Bayes' Rule

The **conditional probability** $P(A|B)$ is the probability that event A occurred given that event B with probability $P(B) > 0$ occurred, which is defined as

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

Notice that, from this definition, we can also say that the probability that event B occurred given that event A with probability $P(A) > 0$ occurred is

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

Now we have two expressions for the probability that both A and B occurred:

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

This now allows us to express **Bayes' rule**, which says that given an event B with probability $P(B) > 0$ has occurred, the probability event A occurred is

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

Our definition of conditional probability also allows us to write the **law of total probability**, which says that if the events $A_1, \ldots, A_n$ are mutually exclusive and collectively exhaustive, then

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

Combining the definition of conditional probability with the law of total probability, we get a more general definition of **Bayes' rule**, which says that if the events $A_1, \ldots, A_n$ are mutually exclusive and collectively exhaustive, then for any event B such that $P(B) > 0$, we have

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

1.3 Independence

Two events A and B are said to be **independent** if the occurrence of one event provides no information about the occurrence of the other. For example, the event of drawing an Ace and the event of rolling a 5 are independent because the fact that I drew an Ace does not impact the probability that I roll a 5. More formally, events A and B are independent if and only if

$$P(A|B) = P(A) \text{ and } P(B|A) = P(B).$$

This is equivalent to saying that A and B are independent if and only if

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

Three events A_1, A_2, and A_3 are **mutually independent** if and only if

1. A_1 and A_2 are independent,
2. A_1 and A_3 are independent,
3. A_2 and A_3 are independent, and
4. $P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3)$.

Extending the definition of mutual independence to n events, we can say that a collection of events $A_1, \ldots, A_n$ are **mutually independent** if and only if for all collections of events $S \subseteq \{1, \ldots, n\}$,

$$P\Big(\bigcap_{i \in S} A_i\Big) = \prod_{i \in S} P(A_i).$$

We also can define conditional independence, which says that A and B are **conditionally independent** given C if and only if

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

Chapter 2

Sequential Experiments

2.1 Counting Methods

Given n total objects, there are $n!$ permutations of those objects. Consider a set of n total objects, and suppose we want to sample $k \leq n$ of those objects. If we choose k out of the n objects without replacement, then there are $(n)_k$ permutations of the objects we choose, where $(n)_k$ is defined as

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

If we are not interested in the order in which the k objects are chosen, then we say there are $\binom{n}{k}$ ways to choose the k objects, where $\binom{n}{k}$ is defined as

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

A partition of a set of n objects is a set of m groups each containing n_i objects such that $n = n_1 + \ldots + n_m$. The number of ways to chose n_i objects of each of the m groups from the set of n total objects is

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

Now consider the set of n total objects, and suppose we want to sample $k \leq n$ of those objects and replace each object as we sample them. Now there are n^k ways to choose an ordered sample of the objects.

2.2 Independent Trials

Independent trials are identical subexperiments in a sequential experiment. Consider an experiment of n independent trials, where the probability of success in each trial is p. Let an outcome of the experiment be defined as a particular

sequence of successes and failures, resulting in a total of n_1 successes and n_2 failures, where $n_1 + n_2 = n$. The probability of one particular outcome is

$$p^{n_1}(1-p)^{n_2}.$$

There are $\binom{n}{n_1} = \binom{n}{n_2}$ possible outcomes that include n_1 successes and $n_2 = n - n_1$ failures. Therefore, the probability of an experiment of n trials, each with probability of success p, resulting in n_1 successes is

$$\binom{n}{n_1} p^{n_1}(1-p)^{n-n_1}.$$

Now we can generalize this result to experiments with more than two possible outcomes. Consider an experiment which has a sample space $\Omega = \{s_1, \ldots, s_m\}$ and the associated probability law $P(s_i) = p_i$. For $n = n_1 + \ldots + n_m$ independent trials, the probability of n_i occurrences of s_i for $i \in \{1, \ldots, m\}$ is given by

$$\binom{n}{n_1, \ldots, n_m} \prod_{i=1}^{m} p_i^{n_i}.$$

2.2.1 Binomial Theorem

As an aside, it is useful to note a related concept, which is sometimes referred to as the **binomial theorem** or the **binomial expansion**. According to the binomial theorem, we can express $(x+y)^n$ for some integer $n \geq 0$ as

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

Part II

Random Variables

Chapter 3

Discrete Random Variables

3.1 Discrete Random Variable

Given an experiment with probability measure P defined on a sample space Ω, a **random variable** is a function that assigns a real number to each outcome. A random variable is a real-valued function $X : \Omega \to \mathbb{R}$ defined such that

$$\{X = x\} = \{\omega \in \Omega : X(\omega) = x\}.$$

In this formulation, the variable x is called a **realization** of the random variable X. A random variable X is discrete if the range of X, which we denote $\mathcal{X}$, is a countable set. This set may contain a finite or an infinite number of values.

3.2 Probability Mass Function

The frequencies with which a random variable takes on different values is described by its **probability mass function (PMF)**. The PMF, $p_X : \mathcal{X} \to [0, 1]$, associated with a discrete random variable X is defined as

$$p_X(x) = P(\{X = x\}) = P(\{\omega \in \Omega : X(\omega) = x\}).$$

From the probability axioms, the PMF satisfies the following properties:

1. $p_X(x) \geq 0, \ \forall x \in \mathcal{X}$
2. $\sum_{x \in \mathcal{X}} p_X(x) = 1$
3. $P(A) = \sum_{x \in A} p_X(x), \ \forall A \in \mathcal{F}$

3.3 Cumulative Distribution Function

The frequencies with which a random variable takes on different values can also be described by its **cummulative distribution function (CDF)**. The CDF,

$F_X : \mathcal{X} \to [0, 1]$, associated with a discrete random variable X is defined as

$$F_X(x) = P(\{X \leq x\}) = P(\{\omega \in \Omega : X(\omega) \leq x\}).$$

If we assume the range is given by $\mathcal{X} = \{x_1, x_2, \ldots\}$, where $x_1 < x_2 < \ldots$, then the CDF of X satisfies the following properties:

1. $F_X(x) = \sum_{x_i \leq x} p_X(x_i), \ \forall x \in \mathcal{X}$
2. $F_X(-\infty) := \lim_{x \to -\infty} F_X(x) = 0$
3. $F_X(\infty) := \lim_{x \to \infty} F_X(x) = 1$
4. $F_X(x') \geq F_X(x), \ \forall x, x' \in \mathcal{X}$ s.t. $x' \geq x$
5. $p_X(x_i) = F_X(x_i) - F_X(x_i - \epsilon), \ \forall x_i \in \mathcal{X}$ and some arbitrarily small $\epsilon > 0$
6. $F_X(x) = F_X(x_i), \ \forall x, x_i, x_{i+1} \in \mathcal{X}$ s.t. $x_i < x < x_{i+1}$
7. $F_X(b) - F_X(a) = P(\{a < X \leq b\}), \ \forall a, b \in \mathcal{X}$

These properties say that the CDF, F_X, starts at zero and increases monotonically to one. Furthermore, there is a discontinuity at each value $x_i \in \mathcal{X}$, and the height of the jump is given by $p_X(x_i)$. Between these discontinuities, the CDF does not change. Finally, the CDF allows us to determine the probability that the random variable X takes on a value between two points.

3.4 Expected Value

For a discrete random variable X with PMF p_X, the **expected value** of X is

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x p_X(x).$$

Given a discrete random variable X with the PMF p_X, $Y = g(X)$ is another discrete random variable with its own PMF, p_Y, which is defined as

$$p_Y(y) = \sum_{x : g(x) = y} p_X(x).$$

This says that $p_Y(y)$ is the sum of probability outcomes $X = x$ for which $Y = y$. For these discrete random variables, the expected value of Y can be defined as

$$\begin{aligned}\mathbb{E}[Y] &= \sum_{y \in \mathcal{Y}} y p_Y(y) = \sum_{y \in \mathcal{Y}} y \left(\sum_{x : g(x) = y} p_X(x) \right) \\ &= \sum_{y \in \mathcal{Y}} \sum_{x : g(x) = y} g(x) p_X(x) = \sum_{x \in \mathcal{X}} g(x) p_X(x).\end{aligned}$$

Now we can see that we compute $\mathbb{E}[Y]$ without needing to compute $p_Y(y)$. This finding is sometimes referred to as the **law of the unconscious statistician**.

3.4.1 Linearity of Expectation

This law of the unconscious statistician leads us to the linearity of expectation, which says that for the discrete random variable X and two constants a and b,

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b.$$

Proof: From the law of the unconscious statistician,

$$\mathbb{E}[aX + b] = \sum_{x \in \mathcal{X}} (ax + b) p_X(x) = a \sum_{x \in \mathcal{X}} x p_X(x) + b \sum_{x \in \mathcal{X}} p_X(x)$$

Using the definition of the expected value for the first term of the sum and the second property of the PMF for the second term, we get the desired result.

3.4.2 Jensen's Inequality

Another useful property of expected value is **Jensen's inequality**, which says that for any discrete random variable X and convex function f,

$$f\big(\mathbb{E}[X]\big) \leq \mathbb{E}\big[f(X)\big].$$

Proof: According to the first order condition for convexity (see convex optimization notes), if f is a convex function, then

$$f(y) \geq f(x) + f'(x)(y - x), \ \forall x, y \in \text{dom} f.$$

If we let $y = X$ and $x = \mathbb{E}[X]$, then this inequality becomes

$$f(X) \geq f(\mathbb{E}[X]) + f'(\mathbb{E}[X])(X - \mathbb{E}[X]).$$

Taking the expectation of both sides and using the linearity of expectation,

$$\begin{aligned} \mathbb{E}[f(X)] &\geq \mathbb{E}\big[f(\mathbb{E}[X]) + f'(\mathbb{E}[X])(X - \mathbb{E}[X])\big] \\ &= f(\mathbb{E}[X]) + f'(\mathbb{E}[X])(\mathbb{E}[X] - \mathbb{E}[X]) \\ &= f(\mathbb{E}[X]). \end{aligned}$$

3.4.3 Tail Sum Formula

The tail sum formula for expectation says that for a non-negative, integer-valued random variable X, the expected value is given by

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} P(\{X \geq k\}).$$

Proof: From the definition of the expected value, we know

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x p_X(x) = \sum_{x \in \mathcal{X}} x P(\{X = x\}).$$

If we assume the X is non-negative and integer-valued, then

$$\begin{aligned}\mathbb{E}[X] &= \sum_{x=1}^{\infty} xP(\{X=x\}) = \sum_{x=1}^{\infty}\sum_{k=1}^{x} P(\{X=x\}) \\ &= \sum_{k=1}^{\infty}\sum_{x=k}^{\infty} P(\{X=x\}) = \sum_{k=1}^{\infty} P(\{X\geq k\}).\end{aligned}$$

3.5 Variance & Standard Deviation

For a discrete random variable X, the **variance** measures the dispersion of its sample values around its expected value, $\mathbb{E}[X]$. The variance of X is defined as

$$\mathrm{Var}(X) = \mathbb{E}\left[\big(X-\mathbb{E}[X]\big)^2\right].$$

We can also express the variance of X using a different formula:

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

Proof: Using the properties of expectation discussed in the previous section, we can show that these expressions are equivalent:

$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}\left[\big(X-\mathbb{E}[X]\big)^2\right] \\ &= \sum_{x\in\mathcal{X}} \big(x-\mathbb{E}[X]\big)^2 p_X(x) \\ &= \sum_{x\in\mathcal{X}} \Big(x^2 - 2x\mathbb{E}[X] + \big(\mathbb{E}[X]\big)^2\Big) p_X(x) \\ &= \sum_{x\in\mathcal{X}} x^2 p_X(x) - 2\mathbb{E}[X]\sum_{x\in\mathcal{X}} x p_X(x) + \big(\mathbb{E}[X]\big)^2 \sum_{x\in\mathcal{X}} p_X(x) \\ &= \mathbb{E}[X^2] - 2\big(\mathbb{E}[X]\big)^2 + \big(\mathbb{E}[X]\big)^2 \\ &= \mathbb{E}[X^2] - \big(\mathbb{E}[X]\big)^2\end{aligned}$$

For the random variable X and two constants a and b, the variance satisfies

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

Proof: We can prove this using the second expression for the variance:

$$\begin{aligned}\mathrm{Var}(aX+b) &= \mathbb{E}[(aX+b)^2] - \big(\mathbb{E}[aX+b]\big)^2 \\ &= \mathbb{E}[a^2X^2 + 2abX + b^2] - \big(\mathbb{E}[aX+b]\big)^2 \\ &= a^2\mathbb{E}[X^2] + 2ab\mathbb{E}[X] + b^2 - \big(a\mathbb{E}[X]+b\big)^2 \\ &= a^2\mathbb{E}[X^2] + 2ab\mathbb{E}[X] + b^2 - a^2(\mathbb{E}[X])^2 - 2ab\mathbb{E}[X] - b^2 \\ &= a^2\mathbb{E}[X^2] - a^2(\mathbb{E}[X])^2 \\ &= a^2\Big(\mathbb{E}[X^2] - (\mathbb{E}[X])^2\Big) = a^2\mathrm{Var}(X)\end{aligned}$$

Another measure of the dispersion of the sample values of a random variable about its expected value is **standard deviation**, which is defined as

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

3.6 Moment Generating Function

Another probability model for the discrete random variable X is the **moment generating function (MGF)**, which is defined for real values of s as

$$\phi_X(s) = \mathbb{E}[e^{sX}] = \sum_{x \in \mathcal{X}} e^{sx} p_X(x).$$

The set of values of s for which $\phi_X(s)$ exists is called the **region of convergence**. MGFs can be used to discuss **moments** of random variables. A discrete random variable X with MGF ϕ_X has the nth moment

$$\mathbb{E}[X^n] = \frac{d^n}{ds^n} \phi_X(s) \bigg|_{s=0}.$$

3.7 Common Discrete Distributions

There are several common types of discrete random variables with well-known distributions. Some common discrete distributions are discussed in this section.

3.7.1 Indicator Random Variable

An indicator random variable X for an event $A \in \mathcal{F}$ is defined as

$$X = \mathbb{1}_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases}.$$

The expected value for this random variable is given by

$$\mathbb{E}[X] = P(A).$$

3.7.2 Uniform Random Variable

The PMF of a random variable $X \sim \text{Uniform}(a, b)$ for $a < b$ is given by

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

Uniform random variables are used if every possible outcome has the same probability. The expected value and variance for this random variable are

$$\mathbb{E}[X] = \frac{a+b}{2} \quad \text{and} \quad \text{Var}(X) = \frac{(b-a)(n-a+2)}{12}.$$

The moment generating function for this random variable is given by

$$\phi_X(s) = \frac{e^{sk} - e^{s(l+1)}}{1 - e^s}.$$

All these expressions can be derived from their definitions using the PMF of X.

3.7.3 Bernoulli Random Variable

The PMF of a random variable $X \sim \text{Bernoulli}(p)$ for $0 \leq p \leq 1$ is given by

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

Bernoulli random variables are used to represent to probability of a success or failure in a single binary experiment, where p is the probability of success. The expected value and variance for this random variable are given by

$$\mathbb{E}[X] = p \quad \text{and} \quad \text{Var}(X) = p(1-p).$$

The moment generating function for this random variable is given by

$$\phi_X(s) = 1 - p + pe^s.$$

All these expressions can be derived from their definitions using the PMF of X.

3.7.4 Binomial Random Variable

The PMF of a random variable $X \sim \text{Binomial}(n, p)$ for $n \in \mathbb{Z}$ and $0 \leq p \leq 1$ is

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

Binomial random variables are used to represent the probability of some number of successes in n binary trials, where p is the probability of success in any given trial. The expected value and variance for this random variable are given by

$$\mathbb{E}[X] = np \quad \text{and} \quad \text{Var}(X) = np(1-p).$$

The moment generating function for this random variable is given by

$$\phi_X(s) = \left(1 - p + pe^s\right)^n.$$

All these expressions can be derived from their definitions using the PMF of X. If $X_1, \ldots, X_n$ are n independent and identically distributed (i.i.d.) (See Section 5.6.1) random variables such that $X_i \sim \text{Bernoulli}(p)$ for $i = \{1, \ldots, n\}$, then

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

If $X \sim \text{Binomial}(n, p)$ and $Y \sim \text{Binomial}(m, p)$ are independent random variables, then by construction, $X + Y \sim \text{Binomial}(n + m, p)$.

3.7.5 Geometric Random Variable

The PMF of a random variable $X \sim \text{Geometric}(p)$ for $0 \leq p \leq 1$ is given by

$$p_X(x) = \begin{cases} p(1-p)^{x-1} & \text{if } x = 1, 2, \dots \\ 0 & \text{otherwise} \end{cases}.$$

Geometric random variables are used to represent to the number of failed trials before the first success, where p is the probability of success in any given trial. The expected value and variance for this random variable are given by

$$\mathbb{E}[X] = \frac{1}{p} \quad \text{and} \quad \text{Var}(X) = \frac{1-p}{p^2}.$$

The moment generating function for this random variable is given by

$$\phi_X(s) = \frac{pe^s}{1-(1-p)e^s}.$$

All these expressions can be derived from their definitions using the PMF of X. If $X_1, \dots, X_n$ are n independent and identically distributed (i.i.d.) (See Section 5.6.1) random variables such that $X_i \sim \text{Bernoulli}(p)$ for $i = \{1, \dots, n\}$, then

$$X = \min\{k \geq 1 : X_k = 1\}$$

The geometric random variable X is **memoryless**, which means

$$P(\{X > t + s | X > s\}) = P(\{X > t\}),\ \forall s, t \geq 0.$$

3.7.6 Poisson Random Variable

The PMF of a random variable $X \sim \text{Poisson}(\lambda)$ for $\lambda > 0$ is given by

$$p_X(x) = \begin{cases} \frac{\lambda^x e^{-\lambda}}{x!} & \text{if } x = 0, 1, \dots \\ 0 & \text{otherwise} \end{cases}.$$

Poisson random variables are often used to represent the number of discrete events that occur within a countably infinite interval of time. The expected value and variance for this random variable are given by

$$\mathbb{E}[X] = \lambda \quad \text{and} \quad \text{Var}(X) = \lambda.$$

These expressions can be derived from their definitions using the PMF of X. By construction of the Poisson distribution, if $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\mu)$ are independent random variables, then the sum of the random variables has the distribution $X + Y \sim \text{Poisson}(\lambda + \mu)$.

3.7.7 Pascal Random Variable

The PMF of a random variable $X \sim \text{Pascal}(k, p)$ is given by

$$p_X(x) = \begin{cases} \binom{x-1}{k-1} p^x (1-p)^{x-k} & \text{if } x = 1, 2, \dots \\ 0 & \text{otherwise} \end{cases}.$$

The expected value and variance for this random variable are given by

$$\mathbb{E}[X] = \frac{k}{p} \quad \text{and} \quad \text{Var}(X) = \frac{k(1-p)}{p^2}.$$

These expressions can be derived from their definitions using the PMF of X.

Chapter 4

Continuous Random Variables

4.1 Continuous Random Variable

Previously, we said that given an experiment with probability measure P defined on a sample space Ω, a random variable is a function that assigns a real number to each outcome in the sample space of the experiment. We defined a random variable as a real-valued function $X : \Omega \to \mathbb{R}$ such that

$$\{X = x\} = \{\omega \in \Omega : X(\omega) = x\}.$$

A random variable X is continuous if the range of X, which we denote $\mathcal{X}$, is not a countable set. Instead, the range consists of one of more intervals of values. Because the probability function is defined over an uncountable range of values, the probability that X takes on a value of exactly x is zero:

$$P(\{X = x\}) = 0.$$

For continuous random variables, we consider the probability that it takes on a value within an interval, rather than taking on an exact value.

4.2 Cumulative Distribution Function

The frequencies with which a continuous random variable takes on different values can be described by its **cumulative distribution function (CDF)**. The CDF, $F_X : \mathcal{X} \to [0, 1]$, associated with a random variable X is defined as

$$F_X(x) = P(\{X \leq x\}) = P(\{\omega \in \Omega : X(\omega) \leq x\}).$$

The CDF of a continuous random variable satisfies the following properties:

1. $F_X(-\infty) := \lim_{x \to -\infty} F_X(x) = 0$

2. $F_X(\infty) := \lim_{x\to\infty} F_X(x) = 1$
3. $F_X(x') \geq F_X(x),\ \forall x, x' \in \mathcal{X}$ s.t $x' \geq x$
4. $F_X(b) - F_X(a) = P(\{a < X \leq b\})$

Note that these properties also hold if X is a discrete random variable. However, if X is discrete, then the CDF has discontinuities that must satisfy additional properties. For a continuous random variable, the CDF is continuous, so it does not need to satisfy these additional properties.

4.3 Probability Density Function

The frequencies with which a continuous random variable takes on different values can also be described by its **probability density function (PDF)**. The PDF, f_X, of a random variable X is defined as the derivative of its CDF:

$$f_X(x) = \frac{d}{dx} F_X(x).$$

The value of $f_X(x)$ indicates the probability that X is near a sample value x and is a good indication of the likely value of observations. The PDF of the continuous random variable X satisfies the following properties:

1. $f_X(x) \geq 0,\ \forall x \in \mathcal{X}$
2. $\int_{\mathcal{X}} f_X(x)dx = 1$
3. $F_X(x) = \int_{-\infty}^{x} f_X(t)dt,\ \forall x \in \mathcal{X}$
4. $\int_a^b f_X(x)dx = P(\{a < X \leq b\}),\ \forall a, b \in \mathcal{X}$

4.4 Expected Value

Similar to the discrete case, for a continuous random variable X with the probability density function (PDF) f_X, the **expected value** of X is defined as

$$\mathbb{E}[X] = \int_{\mathcal{X}} x f_X(x)dx.$$

Given a continuous random variable X with the PDF f_X, $Y = g(X)$ is another random variable, but it is not necessarily continuous, so it may not have a well-defined PDF. However, similar to as we showed in the discrete case, the expected value of Y can be defined in terms of the distribution of X as

$$\mathbb{E}[Y] = \int_{\mathcal{X}} g(x) f_X(x).$$

As in the discrete case, we can compute $\mathbb{E}[Y]$ without needing to compute $f_Y(y)$, which is called the **law of the unconscious statistician**.

4.4.1 Linearity of Expectation

In the same way as we saw in the discrete case, the law of the unconscious statistician leads us to the linearity of expectation, which says that for the continuous random variable X and two constants a and b,

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b.$$

Proof: From the law of the unconscious statistician,

$$\mathbb{E}[aX + b] = \int_{\mathcal{X}} (ax + b) f_X(x) dx = a \int_{\mathcal{X}} x f_X(x) dx + b \int_{\mathcal{X}} f_X(x) dx$$

Using the definition of the expected value for the first term of the sum and the second property of the PDF for the second term, we get the desired result.

4.4.2 Jensen's Inequality

Jensen's inequality also holds for continuous random variables and says that for any continuous random variable X and convex function f,

$$f\big(\mathbb{E}[X]\big) \leq \mathbb{E}\big[f(X)\big].$$

The proof for this inequality is exactly the same as in the discrete case.

4.4.3 Tail Sum Formula

The tail sum formula for expectation also applies to continuous random variables and says that for a non-negative random variable X, the expected value is

$$\mathbb{E}[X] = \int_0^\infty P(\{X \geq t\})\ dt = \int_0^\infty \big(1 - F_X(t)\big)\ dt.$$

More more information about the tail sim formula, see: `https://stat.uiowa.edu/sites/stat.uiowa.edu/files/cae/Lo_Expectation.pdf`

4.5 Variance & Standard Deviation

For a continuous random variable X, its **variance** measures the dispersion of sample values of X around its expected value $\mathbb{E}[X]$. The variance of X is

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

As in the discrete case, we can also express the variance as

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

Proof: Using the properties of expectation discussed in the previous section, we can show that these expressions are equivalent:

$$\begin{aligned}
\text{Var}(X) &= \mathbb{E}\left[\big(X - \mathbb{E}[X]\big)^2\right] \\
&= \int_{\mathcal{X}} \big(x - \mathbb{E}[X]\big)^2 f_X(x)dx \\
&= \int_{\mathcal{X}} \Big(x^2 - 2x\mathbb{E}[X] + \big(\mathbb{E}[X]\big)^2\Big) f_X(x)dx \\
&= \int_{\mathcal{X}} x^2 f_X(x) - 2\mathbb{E}[X]\int_{\mathcal{X}} x f_X(x) + \big(\mathbb{E}[X]\big)^2 \int_{\mathcal{X}} f_X(x) \\
&= \mathbb{E}[X^2] - 2\big(\mathbb{E}[X]\big)^2 + \big(\mathbb{E}[X]\big)^2 \\
&= \mathbb{E}[X^2] - \big(\mathbb{E}[X]\big)^2
\end{aligned}$$

An important property of variance, which we showed in the discrete case, is that for the continuous random variable X and two constants a and b,

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

The proof for the statement is the same as was shown for the discrete case. Another measure of the dispersion of the sample values of a random variable about its expected value is standard deviation, which is defined as

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

4.6 Moment Generating Function

Another probability model for the continuous random variable X is the **moment generating function (MGF)**, which is defined for real values of s as

$$\phi_X(s) = \mathbb{E}[e^{sX}] = \int_{\mathcal{X}} e^{sx} f_X(x)dx.$$

The set of values of s for which $\phi_X(s)$ exists is called the **region of convergence**. A continuous random variable X with MGF $\phi_X(s)$ has the nth moment

$$\mathbb{E}[X^n] = \frac{d^n}{ds^n}\phi_X(s)\bigg|_{s=0}.$$

4.7 Common Continuous Distributions

There are several common types of continuous random variables with well-known distributions. Some common continuous distributions are discussed here.

4.7.1 Uniform Random Variable

The PDF of a random variable $X \sim \text{Uniform}(a, b)$ for $a < b$ is given by

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

Uniform random variables are used to model situations in which there is an equal chance of finding an outcome x in any given interval on (a, b). The CDF for the uniform random variable X is given by

$$F_X(x) = \begin{cases} 0 & \text{if } x \leq a \\ \frac{x-a}{b-a} & \text{if } a < x \leq b \\ 1 & \text{if } x > b \end{cases}.$$

The expected value and variance for this random variable are

$$\mathbb{E}[X] = \frac{a+b}{2} \quad \text{and} \quad \text{Var}(X) = \frac{(b-a)^2}{12}.$$

The moment generating function for this random variable is given by

$$\phi_X(s) = \frac{e^{bs} - e^{as}}{s(b-a)}.$$

All these expressions can be derived from their definitions using the PDF of X.

4.7.2 Exponential Random Variable

The PDF of a random variable $X \sim \text{Exponential}(\lambda)$ for $\lambda > 0$ is given by

$$f_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ \lambda e^{-\lambda x} & \text{if } x \geq 0 \end{cases}.$$

Exponential random variables are used to model the amount of time that passes before an event occurs. The CDF for the exponential random variable X is

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - e^{-\lambda x} & \text{if } x \geq 0 \end{cases}.$$

The expected value and variance for this random variable are

$$\mathbb{E}[X] = \frac{1}{\lambda} \quad \text{and} \quad \text{Var}(X) = \frac{1}{\lambda^2}.$$

The moment generating function for this random variable is given by

$$\phi_X(s) = \frac{\lambda}{\lambda - s}.$$

All these expressions can be derived from their definitions using the PDF of X. The exponential random variable X is **memoryless**, which means

$$P(\{X > t + s | X > s\}) = P(\{X > t\}),\ \forall s, t \geq 0.$$

4.7.3 Erlang Random Variable

The PDF of a random variable $X \sim \text{Erlang}(n, \lambda)$ for $\lambda > 0$ and integer $n \geq 1$ is

$$f_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{\lambda^n x^{n-1} e^{-\lambda x}}{(n-1)!} & \text{if } x \geq 0 \end{cases}.$$

The expected value and variance for this random variable are given by

$$\mathbb{E}[X] = \frac{n}{\lambda} \quad \text{and} \quad \text{Var}(X) = \frac{n}{\lambda^2}.$$

These expressions can be derived from their definitions using the PDF of X. An Erlang random variable with parameters $\lambda > 0$ and $n = 1$ is an exponential random variable with parameter λ.

4.7.4 Gaussian Random Variable

The PDF of a random variable $X \sim N(\mu, \sigma^2)$ for real number μ and $\sigma > 0$ is

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

The graph of $f_X(x)$ has a bell shape and is centered at μ. The parameter σ reflects the width of the bell such that smaller values of σ correspond to a narrow bell shape with a high peak and larger values of σ correspond to a wide bell with a low peak. The height of the peak is given by $1/\sqrt{2\pi\sigma^2}$. The expected value and variance for this random variable are given by

$$\mathbb{E}[X] = \mu \quad \text{and} \quad \text{Var}(X) = \sigma^2.$$

The moment generating function for this random variable is given by

$$\phi_X(s) = \exp\Big(s\mu + \frac{1}{2}s^2\sigma^2\Big).$$

All these expressions can be derived from their definitions using the PDF of X.

4.7.5 Standard Normal Random Variable

The standard normal random variable $Z \sim N(0, 1)$ is the Gaussian random variable with zero mean and unit variance, which has the PDF

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

To prove that this is actually a valid PDF, first notice that this function is non-negative for all values of z. Now we need to show $\int_{-\infty}^{\infty} f_Z(z)dz = 1$. To do

so, we will consider the square of the given integral:

$$\begin{aligned}
\left(\int_{-\infty}^{\infty} f_Z(z)dz\right)^2 &= \left(\int_{-\infty}^{\infty} f_Z(u)du\right)\left(\int_{-\infty}^{\infty} f_Z(v)dv\right) \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_Z(u)f_Z(v)dudv \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left(\frac{1}{\sqrt{2\pi}}e^{-u^2/2}\right)\left(\frac{1}{\sqrt{2\pi}}e^{-v^2/2}\right)dudv \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{1}{2\pi}e^{-(u^2+v^2)/2}dudv \\
&= \int_{0}^{2\pi}\int_{0}^{\infty} \frac{1}{2\pi}e^{-r^2/2}rdrd\theta \\
&= \int_{0}^{2\pi} \frac{1}{2\pi}d\theta \int_{0}^{\infty} e^{-r^2/2}rdr \\
&= \frac{1}{2\pi}(2\pi)\int_{0}^{\infty} e^{-t}dt \\
&= 1 \checkmark
\end{aligned}$$

The CDF for the standard normal random variable Z is given by

$$\phi(z) := F_Z(z) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-u^2/2}du.$$

The CDF for Z satisfies an important symmetrical property:

$$\phi(-z) = 1 - \phi(z).$$

The standard normal complementary CDF is given by

$$Q(z) = 1 - \phi(z) = \frac{1}{\sqrt{2\pi}}\int_{z}^{\infty} e^{-u^2/2}du.$$

To transform a Gaussian random variable $X \sim N(\mu, \sigma^2)$ into a standard normal random variable $Z \sim N(0, 1)$, we can define Z as

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

We can then define the probability distributions of the Gaussian random variable X with mean μ and standard deviation σ in terms of the standard normal CDF:

$$F_X(x) = \phi\left(\frac{x-\mu}{\sigma}\right) \quad \text{and} \quad P(\{a < X \leq b\}) = \phi\left(\frac{b-\mu}{\sigma}\right) - \phi\left(\frac{a-\mu}{\sigma}\right).$$

Part III

Multiple Random Variables

Chapter 5

Multiple Random Variables

5.1 Joint Probability Mass Function

5.1.1 Two Random Variables

Let X and Y be two discrete random variables defined on the same probability space, $(\Omega, \mathcal{F}, P)$. Suppose the range of X and Y are $\mathcal{X}$ and $\mathcal{Y}$ respectively. Their **joint probability mass function (PMF)** describes the frequencies of their joint outcomes. The joint PMF of X and Y is defined as

$$\begin{aligned} p_{XY}(x,y) &= P(\{X=x, Y=y\}) \\ &= P(\{\omega \in \Omega : X(\omega)=x, Y(\omega)=y\}) \\ &= P(\{\omega \in \Omega : X(\omega)=x\} \cap \{\omega \in \Omega : Y(\omega)=y\}). \end{aligned}$$

The joint probability mass function satisfies the following properties:

1. $p_{XY}(x,y) \geq 0, \ \forall x \in \mathcal{X}, y \in \mathcal{Y}$
2. $\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p_{XY}(x,y) = 1$
3. $P(A) = \sum_{(x,y)\in A} p_{XY}(x,y), \ \forall A \in \mathcal{F}$

For random variables X and Y with joint PMF p_{XY}, the **marginal PMFs** are probability models for the individual random variables and are defined as

$$p_X(x) = \sum_{y\in\mathcal{Y}} p_{XY}(x,y) \quad \text{and} \quad p_Y(y) = \sum_{x\in\mathcal{X}} p_{XY}(x,y).$$

5.1.2 Multiple Random Variables

We can also expand the definition of the joint probability mass function (PMF) to n random variables. Let $X_1, \ldots, X_n$ be n discrete random variables defined

on the same probability space, $(\Omega, \mathcal{F}, P)$. Suppose the range of X_i is $\mathcal{X}_i$ for $i = 1, \ldots, n$. For the random variables $X_1, \ldots, X_n$, the joint PMF is

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

For n random variables, the joint PMF satisfies the following properties:

1. $p_{X_1 \ldots X_n}(x_1, \ldots, x_n) \geq 0, \ \forall x_1 \in \mathcal{X}_1, \ldots, x_n \in \mathcal{X}_n$
2. $\sum_{x_1 \in \mathcal{X}_1} \cdots \sum_{x_n \in \mathcal{X}_n} p_{X_1 \ldots X_n}(x_1, \ldots, x_n) = 1$
3. $P(A) = \sum_{(x_1, \ldots, x_n) \in A} p_{X_1 \ldots X_n}(x_1, \ldots, x_n), \ \forall A \in \mathcal{F}$

Consider four random variables X_1, X_2, X_3, and X_4 with joint PMF $p_{X_1 X_2 X_3 X_4}$. The marginal PMFs can be computed in the following way:

$$p_{X_2 X_3 X_4}(x_2, x_3, x_4) = \sum_{x_1 \in \mathcal{X}_1} p_{X_1 X_2 X_3 X_4}(x_1, x_2, x_3, x_4)$$

$$p_{X_1 X_4}(x_1, x_4) = \sum_{x_2 \in \mathcal{X}_2} \sum_{x_3 \in \mathcal{X}_3} p_{X_1 X_2 X_3 X_4}(x_1, x_2, x_3, x_4)$$

5.2 Joint Cumulative Distribution Function

5.2.1 Two Random Variables

Let X and Y be two random variables defined on the same probability space, $(\Omega, \mathcal{F}, P)$. Suppose the range of X and Y are $\mathcal{X}$ and $\mathcal{Y}$ respectively. Their **joint cumulative distribution function (CDF)** describes the frequencies of their joint outcomes. The joint CDF of X and Y is defined as

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

The joint cumulative distribution function (CDF) satisfies several properties:

1. $0 \leq F_{XY}(x, y) \leq 1, \ \forall x \in \mathcal{X}, y \in \mathcal{Y}$
2. $F_{XY}(\infty, \infty) := \lim_{x \to \infty} \lim_{y \to \infty} F_{XY}(x, y) = 1$
3. $F_{XY}(x, -\infty) := \lim_{y \to -\infty} F_{XY}(x, y) = 0, \ \forall x \in \mathcal{X}$
4. $F_{XY}(-\infty, y) := \lim_{x \to -\infty} F_{XY}(x, y) = 0, \ \forall y \in \mathcal{Y}$
5. $F_X(x) = F_{XY}(x, \infty) := \lim_{y \to \infty} F_{XY}(x, y), \ \forall x \in \mathcal{X}$
6. $F_Y(y) = F_{XY}(\infty, y) := \lim_{x \to \infty} F_{XY}(x, y), \ \forall y \in \mathcal{Y}$
7. $F_{XY}(x_1, y_1) \leq F_{XY}(x_2, y_2), \ \forall x_1, x_2 \in \mathcal{X}, y_1, y_2 \in \mathcal{Y} : x_1 \leq x_2, y_1 \leq y_2$
8. $P(\{x_1 < X \leq x_2, y_1 < Y \leq y_2\}) =$
   $F_{XY}(x_2, y_2) - F_{XY}(x_2, y_1) - F_{XY}(x_1, y_2) + F_{XY}(x_1, y_1),$
   $\forall x_1, x_2 \in \mathcal{X}, y_1, y_2 \in \mathcal{Y}$ s.t. $x_1 \leq x_2$ and $y_1 \leq y_2$

5.2.2 Multiple Random Variables

We can also expand the definition of the joint cumulative distribution function (CDF) to n random variables. Let $X_1, \ldots, X_n$ be n random variables defined on the same probability space, $(\Omega, \mathcal{F}, P)$. The joint CDF is defined as

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

The joint CDF satisfies similar properties as in the case of two random variables.

5.3 Joint Probability Density Function

5.3.1 Two Random Variables

Let X and Y be two continuous random variables defined on the same probability space, $(\Omega, \mathcal{F}, P)$. Suppose the range of X and Y are $\mathcal{X}$ and $\mathcal{Y}$ respectively. The **joint probability density function (PDF)** of the continuous random variables X and Y is defined in terms of their joint CDF as

$$f_{XY}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{XY}(x, y).$$

The joint PDF of random variables X and Y satisfies the following properties:

1. $f_{XY}(x, y) \geq 0, \ \forall x \in \mathcal{X}, y \in \mathcal{Y}$
2. $\int_{\mathcal{X}} \int_{\mathcal{Y}} f_{XY}(x, y) dx dy = 1$
3. $F_{XY}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{XY}(u, v) du dv$
4. $P(A) = \iint_A f_{XY}(x, y) dx dy, \ \forall A \in \mathcal{F}$

For random variables X and Y with joint PDF f_{XY}, the **marginal PDFs** are probability models for the individual random variables and are defined as

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y) dy \quad \text{and} \quad f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x, y) dx.$$

5.3.2 Multiple Random Variables

We can also expand the definition of the joint probability density function (PDF) to n random variables. Let $X_1, \ldots, X_n$ be n continuous random variables defined on the same probability space, $(\Omega, \mathcal{F}, P)$. Suppose the range of X_i is $\mathcal{X}_i$ for $i = 1, \ldots, n$. For the random variables $X_1, \ldots, X_n$, the joint PDF is

$$f_{X_1 \ldots X_n}(x_1, \ldots, x_n) = \frac{\partial^n}{\partial x_1 \ldots \partial x_n} F_{X_1 \ldots X_n}(x_1, \ldots, x_n)$$

For n random variables, the joint PDF satisfies the following properties:

1. $f_{X_1 \ldots X_n}(x_1, \ldots, x_n) \geq 0, \ \forall x_1 \in \mathcal{X}_1, \ldots, x_n \in \mathcal{X}_n$
2. $\int_{\mathcal{X}_1} \ldots \int_{\mathcal{X}_n} f_{X_1 \ldots X_n}(x_1, \ldots, x_n) dx_1 \ldots dx_n = 1$
3. $F_{X_1 \ldots X_n}(x_1, \ldots, x_n) = \int_{-\infty}^{x_1} \ldots \int_{-\infty}^{x_n} f_{X_1 \ldots X_n}(u_1, \ldots, u_n) du_1 \ldots du_n$
4. $P(A) = \int \ldots \int_A f_{X_1 \ldots X_n}(x_1, \ldots, x_n) dx_1 \ldots dx_n, \ \forall A \in \mathcal{F}$

Consider four random variables X_1, X_2, X_3, and X_4 with joint PDF is $f_{X_1 X_2 X_3 X_4}$. The marginal PDFs can be computed in the following way:

$$f_{X_2 X_3 X_4}(x_2, x_3, x_4) = \int_{\mathcal{X}_1} f_{X_1 X_2 X_3 X_4}(x_1, x_2, x_3, x_4) dx_1$$

$$f_{X_1 X_4}(x_1, x_4) = \int_{\mathcal{X}_2} \int_{\mathcal{X}_3} f_{X_1 X_2 X_3 X_4}(x_1, x_2, x_3, x_4) dx_2 dx_3$$

5.4 Expected Value

5.4.1 Discrete Random Variables

Given a two discrete random variables X and Y with joint PMF p_{XY}, we can define a new random variable $W = g(X, Y)$. The expected value of W is

$$\mathbb{E}[W] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} g(x, y) p_{XY}(x, y).$$

Using this definition of the expected value, we can show that the expectation of a sum of two discrete random variables is the sum of its expected values:

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y].$$

Proof: We can directly prove this property for two discrete random variables:

$$\begin{aligned}
\mathbb{E}[X + Y] &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} (x + y) p_{XY}(x, y) \\
&= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} x p_{XY}(x, y) + \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} y p_{XY}(x, y) \\
&= \sum_{x \in \mathcal{X}} x \left(\sum_{y \in \mathcal{Y}} p_{XY}(x, y) \right) + \sum_{y \in \mathcal{Y}} y \left(\sum_{x \in \mathcal{X}} p_{XY}(x, y) \right) \\
&= \sum_{x \in \mathcal{X}} x p_X(x) + \sum_{y \in \mathcal{Y}} y p_Y(y) \\
&= \mathbb{E}[X] + \mathbb{E}[Y]
\end{aligned}$$

5.4.2 Continuous Random Variables

Given a two continuous random variables X and Y with joint PDF f_{XY}, we can define a new random variable $W = g(X, Y)$. The expected value of W is

$$\mathbb{E}[W] = \int_{\mathcal{X}} \int_{\mathcal{Y}} g(x, y) f_{XY}(x, y) dx dy.$$

As for the discrete case, the expectation of a sum of two continuous random variables is the sum of its expected values:

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y].$$

Proof: We can directly prove this for two continuous random variables:

$$\begin{aligned}
\mathbb{E}[X + Y] &= \int_{\mathcal{X}} \int_{\mathcal{Y}} (x + y) f_{XY}(x, y) dx dy \\
&= \int_{\mathcal{X}} \int_{\mathcal{Y}} x f_{XY}(x, y) dx dy + \int_{\mathcal{X}} \int_{\mathcal{Y}} y f_{XY}(x, y) dx dy \\
&= \int_{\mathcal{X}} x \left(\int_{\mathcal{Y}} f_{XY}(x, y) dy \right) dx + \int_{\mathcal{Y}} y \left(\int_{\mathcal{X}} f_{XY}(x, y) dx \right) dy \\
&= \int_{\mathcal{X}} x f_X(x) dx + \int_{\mathcal{Y}} y f_Y(y) dy \\
&= \mathbb{E}[X] + \mathbb{E}[Y]
\end{aligned}$$

5.5 Variance, Covariance, & Correlation

Previously, we considered the variance as a measure of the dispersion of sample values for a single random variable. We can also consider the variance of multiple random variables. This leads to two related metrics: covariance and correlation.

5.5.1 Variance

For any two random variables X and Y, the variance of their sum can be expressed in terms of the variances of the individual variables as

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathbb{E}\left[\big(X - \mathbb{E}[X]\big) \big(Y - \mathbb{E}[Y]\big) \right].$$

Proof: We can use the definition of the variance to show this property:

$$\begin{aligned}
\mathrm{Var}(X+Y) &= \mathbb{E}\left[\big(X+Y-\mathbb{E}[X+Y]\big)^2\right] \\
&= \mathbb{E}\left[\big(X+Y-\mathbb{E}[X]-\mathbb{E}[Y]\big)^2\right] \\
&= \mathbb{E}\Big[X^2+Y^2+\big(\mathbb{E}[X]\big)^2+\big(\mathbb{E}[X]\big)^2+2XY-2X\mathbb{E}[X]-2Y\mathbb{E}[X]-2X\mathbb{E}[Y]-2Y\mathbb{E}[Y]+2\mathbb{E}[X]\mathbb{E}[Y]\Big] \\
&= \mathbb{E}\left[\big(X-\mathbb{E}[X]\big)^2+\big(Y-\mathbb{E}[Y]\big)^2+2\big(X-\mathbb{E}[X]\big)\big(Y-\mathbb{E}[Y]\big)\right] \\
&= \mathbb{E}\left[\big(X-\mathbb{E}[X]\big)^2\right]+\mathbb{E}\left[\big(Y-\mathbb{E}[Y]\big)^2\right]+2\mathbb{E}\left[\big(X-\mathbb{E}[X]\big)\big(Y-\mathbb{E}[Y]\big)\right] \\
&= \mathrm{Var}(X)+\mathrm{Var}(Y)+2\mathbb{E}\left[\big(X-\mathbb{E}[X]\big)\big(Y-\mathbb{E}[Y]\big)\right]
\end{aligned}$$

5.5.2 Covariance

The last term in the equation above is called the **covariance** and is denoted:

$$\sigma_{XY} := \mathrm{Cov}(X,Y) = \mathbb{E}\left[\big(X-\mathbb{E}[X]\big)\big(Y-\mathbb{E}[Y]\big)\right].$$

The covariance describes how the pair of random variables, X and Y, vary together. If the covariance is positive, then as X increases, Y generally increases as well. If the covariance is negative, then as X increases, Y generally decreases. If the covariance is zero, then X and Y are said to be **uncorrelated**. The covariance satisfies the following useful linearity property:

$$\begin{aligned}
&\mathrm{Cov}\left(\sum_{j=1}^{n} A_{ij}X_j+b_i,\ \sum_{l=1}^{n} A_{kl}X_l+b_k\right) \\
&\qquad = \mathrm{Cov}\left(\sum_{j=1}^{n} A_{ij}X_j,\ \sum_{l=1}^{n} A_{kl}X_l\right) \\
&\qquad = \sum_{j=1}^{n}\sum_{l=1}^{n} A_{ij}A_{kl}\mathrm{Cov}(X_j,X_l) \\
&\qquad = \sum_{j=1}^{n} A_{ij}A_{kj}\mathrm{Var}(X_j)+\sum_{1\le j<l\le n} A_{ij}A_{kl}\mathrm{Cov}(X_j,X_l)
\end{aligned}$$

5.5.3 Correlation Coefficient

The **correlation coefficient** is a normalized version of the covariance, which indicates the relationship between two random variables, regardless of the measurement units. While the unit of the covariance is the product of the units of X and Y, the correlation coefficient is dimensionless and is unaffected by scale changes. The correlation coefficient for two random variables X and Y is

$$\rho_{XY} := \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}.$$

The correlation coefficient, ρ_{XY}, is restricted to values between -1 and 1. If ρ_{XY} is positive, there is a positive correlation between X and Y. Similarly, if it is negative, then there is a negative correlation. If the correlation coefficient is zero, then there is no correlation. If the magnitude of ρ_{XY} is close to zero, then the random variables are only weakly correlated. If the magnitude of ρ_{XY} is close to one, then the random variables are highly correlated. If it is equal to one, then there is a linear relationship between the two variables.

5.5.4 Correlation

Another metric used to describe the relationship between two random variables is **correlation**. The correlation between random variables X and Y is

$$r_{XY} := \mathbb{E}[XY] = \text{Cov}(X, Y) + \mathbb{E}[X]\mathbb{E}[Y].$$

If the correlation between two random variables, X and Y, is zero, then X and Y are said to be **orthogonal**.

5.6 Independent Random Variables

5.6.1 Discrete Random Variables

Two discrete random variables X and Y, with corresponding ranges $\mathcal{X}$ and $\mathcal{Y}$, are considered to be **independent** if and only if

$$p_{XY}(x, y) = p_X(x)p_Y(y), \ \forall x \in \mathcal{X}, y \in \mathcal{Y}.$$

Similarly, n discrete random variables $X_1, \ldots, X_n$, with corresponding ranges $\mathcal{X}_1, \ldots, \mathcal{X}_n$, are considered to be independent if and only if

$$p_{X_1 \ldots X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} p_{X_i}(x_i), \ \forall x_i \in \mathcal{X}_i, i = 1, \ldots, n.$$

These variables are said to be **independent and identically distributed (i.i.d.)** if and only if (1) the random variables are independent and (2) all of the random variables have the same PMF, which we denote p_X, meaning

$$p_{X_1 \ldots X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} p_X(x_i).$$

5.6.2 Continuous Random Variables

Two continuous random variables X and Y, with corresponding ranges $\mathcal{X}$ and $\mathcal{Y}$, are considered to be **independent** if and only if

$$f_{XY}(x, y) = f_X(x)f_Y(y), \forall x \in \mathcal{X}, y \in \mathcal{Y}.$$

Similarly, n continuous random variables $X_1, \ldots, X_n$, with corresponding ranges $\mathcal{X}_1, \ldots, \mathcal{X}_n$, are considered to be independent if and only if

$$f_{X_1 \ldots X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i), \ \forall x_i \in \mathcal{X}_i, i = 1, \ldots, n.$$

These variables are said to be **independent and identically distributed (i.i.d.)** if and only if (1) the random variables are independent and (2) all of the random variables have the same PDF, which we denote f_X, meaning

$$f_{X_1 \ldots X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_X(x_i)$$

5.6.3 Independence Properties

For two independent variables X and Y, the following properties hold:

1. $\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)]$
2. $r_{XY} = \mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$
3. $\text{Cov}(X, Y) = 0$
4. $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$

Notice that all independent random variables are uncorrelated, but they are not necessarily orthogonal. Also note that while independence implies that two random variables are uncorrelated, two uncorrelated random variables are not necessarily independent.

5.7 Bivariate Gaussian Random Variables

One special class of a pair of random variables is called bivariate Gaussian random variables. Let X be a Gaussian random variable with mean μ_X and standard deviation σ_X, and let Y be a Gaussian random variable with mean μ_Y and standard deviation σ_Y. Assume that the correlation coefficient between X and Y is given by ρ_{XY}. If X and Y form a pair of bivariate Gaussian random variables, then their joint PDF can be expressed as the following:

$$f_{XY}(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho_{XY}^2}} \exp\left(-\frac{\left(\frac{x-\mu_X}{\sigma_X}\right)^2 + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2 - \frac{2\rho_{XY}(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}}{2(1-\rho_{XY}^2)} \right).$$

Previously, we said two independent variables are uncorrelated but two uncorrelated random variables are not necessarily independent. For bivariate Gaussian random variables, they are uncorrelated if and only if they are independent.

Another important property of bivariate Gaussian random variables is that a pair of linear combinations of bivariate Gaussian random variables forms another pair of bivariate Gaussian random variables. If X and Y form a pair of bivariate Gaussian random variables with parameters $(\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho_{XY})$, and the random variables W_1 and W_2 are given by the linearly independent equations $W_1 = a_1X + b_1Y$ and $W_2 = a_2X + b_2Y$, then W_1 and W_2 form a pair of bivariate Gaussian random variables such that

$$(1)\ \mathbb{E}[W_i] = a_i\mu_X + b_i\mu_Y,$$

$$(2)\ \mathrm{Var}(W_i) = a_i^2\sigma_X^2 + b_i^2\sigma_Y^2 + 2a_ib_i\rho_{XY}\sigma_X\sigma_Y, \text{ and}$$

$$(3)\ \mathrm{Cov}(W_1, W_2) = a_1a_2\sigma_X^2 + b_1b_2\sigma_Y^2 + (a_1b_2 + a_2b_1)\rho_{XY}\sigma_X\sigma_Y$$

Chapter 6

Derived Random Variables

6.1 Derived Discrete Random Variable

6.1.1 Function of One Random Variable

Given a discrete random variable X with PMF p_X and some function g, the derived random variable $Y = g(X)$ has the PMF given by

$$p_Y(y) = \sum_{x:g(x)=y} p_X(x).$$

If g is an invertible function, then we can express the PMF as

$$p_Y(y) = p_X(g^{-1}(y)).$$

6.1.2 Function of Two Random Variables

We can extend the definition given previously for functions of two random variables. Given discrete random variables X and Y with PMFs p_X and p_Y respectively, the derived random variable $W = g(X, Y)$ has the PMF given by

$$p_W(w) = \sum_{(x,y):g(x,y)=w} p_{XY}(x, y).$$

6.1.3 Sums of Independent Random Variables

If X and Y are discrete random variables, then their sum is another discrete random variable, $W = X + Y$. If X and Y are independent with PMFs p_X and p_Y respectively, then the PMF of W is the convolution of p_X and p_Y:

$$p_W(w) = (p_X * p_Y)(w) = \sum_{k=-\infty}^{\infty} p_X(k)p_Y(w - k).$$

6.2 Derived Continuous Random Variable

6.2.1 Function of One Random Variable

Given a continuous random variable X with CDF F_X and some function g, the derived random variable $Y = g(X)$ has the CDF given by

$$F_Y(y) = P(\{Y \leq y\}) = P(\{g(X) \leq y\}).$$

If g is an invertible function, then we can express the CDF as

$$F_Y(y) = P(\{X \leq g^{-1}(y)\}) = F_X(g^{-1}(y)).$$

Given the CDF of the derived random variable Y, we can find its probability density function (PDF) by computing the derivative of the CDF:

$$f_Y(y) = \frac{d}{dy} F_Y(y).$$

If g is a linear function with slope a and intercept b, then $Y = aX + b$. For this special case, we can use the following identities to derive the CDF and PDF Y:

$$F_Y(y) = F_X\Big(\frac{y-b}{a}\Big) \quad \text{and} \quad f_Y(y) = \frac{1}{a} f_X\Big(\frac{y-b}{a}\Big).$$

6.2.2 Function of Two Random Variables

We can extend the definition given previously for functions of two random variables. Given continuous random variables X and Y with CDFs F_X and F_Y respectively, the derived random variable $W = g(X, Y)$ has the CDF given by

$$F_W(w) = \iint_{(x,y):g(x,y)\leq w} f_{XY}(x, y) dx dy.$$

Again, the probability density function (PDF) of W is then given by

$$f_W(w) = \frac{d}{dw} F_W(w).$$

6.2.3 Sums of Independent Random Variables

If X and Y are continuous random variables, then their sum is another continuous random variable, $W = X + Y$. If X and Y are independent with PDFs f_X and f_Y respectively, then the PDF of W is the convolution of f_X and f_Y:

$$f_W(w) = (f_X * f_Y)(w) = \int_{-\infty}^{\infty} f_X(\tau) f_Y(w - \tau) d\tau$$

Chapter 7

Conditional Probability Models

7.1 Conditioning by an Event

7.1.1 Conditional Probability Mass Function

For a discrete random variable X with PMF p_X and an event A with probability $P(A) > 0$, the **conditional probability mass function (PMF)** is defined as

$$p_{X|A}(x) = P(\{X = x|A\}) = \begin{cases} \frac{p_X(x)}{P(A)} & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases}.$$

The conditional PMF satisfies the following properties:

1. $p_{X|A}(x) \geq 0, \ \forall x \in A$
2. $\sum_{x \in A} p_{X|A}(x) = 1$
3. $P(B|A) = \sum_{x \in B} p_{X|A}(x)$

For a discrete random variable X resulting from an experiment with partition $A_1, \ldots, A_m$, we can express the PMF of X in terms of its conditional PMF as

$$p_X(x) = \sum_{i=1}^{m} p_{X|A_i}(x) P(A_i).$$

For two discrete random variables X and Y with the joint PMF p_{XY} and an event A with probability $P(A) > 0$, the **conditional joint PMF** is given by

$$p_{XY|A}(x, y) = P(\{X = x, Y = y|A\}) = \begin{cases} \frac{p_{XY}(x,y)}{P(A)} & \text{if } (x, y) \in A \\ 0 & \text{otherwise} \end{cases}.$$

7.1.2 Conditional Cumulative Distribution Function

For a random variable X with CDF F_X and event A with probability $P(A) > 0$, the **conditional cumulative distribution function (CDF)** is defined as

$$F_{X|A}(x) = P(\{X \leq x|A\}) = \begin{cases} \frac{F_X(x)}{P(A)} & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases}.$$

For two random variables X and Y with joint CDF F_{XY} and an event A with probability $P(A) > 0$, the **conditional joint CDF** is given by

$$F_{XY|A}(x,y) = P(\{X \leq x, Y \leq y|A\}) = \begin{cases} \frac{F_{XY}(x,y)}{P(A)} & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases}.$$

7.1.3 Conditional Probability Density Function

For a continuous random variable X with PDF f_X and conditional CDF $F_{X|A}$ and an event A with probability $P(A) > 0$, the **conditional probability density function (PDF)** can be derived from the conditional CDF as

$$f_{X|A}(x) = \frac{d}{dx}F_{X|A}(x) = \begin{cases} \frac{f_X(x)}{P(A)} & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases}.$$

The conditional PDF satisfies the following properties:

1. $f_{X|A}(x) \geq 0, \ \forall x \in A$
2. $\int_A f_{X|A}(x)dx = 1$
3. $P(B|A) = \int_B p_{X|A}(x)dx$

For a continuous random variable X resulting from an experiment with partition $A_1, \ldots, A_m$, we can express the PDF of X in terms of its conditional PDF as

$$f_X(x) = \sum_{i=1}^{m} f_{X|A_i}(x)P(A_i).$$

For two continuous random variables X and Y with joint PDF f_{XY} and conditional joint CDF $F_{XY|A}$ and an event A with probability $P(A) > 0$, the **conditional joint PDF** is given by

$$f_{XY|A}(x,y) = \frac{\partial^2}{\partial x \partial y}F_{XY|A}(x,y) = \begin{cases} \frac{f_{XY}(x,y)}{P(A)} & \text{if } (x,y) \in A \\ 0 & \text{otherwise} \end{cases}.$$

7.1.4 Conditional Expected Value

For a *discrete* random variable X and an event A with probability $P(A) > 0$, the **conditional expected value** of X is defined as

$$\mathbb{E}[X|A] = \sum_{x \in \mathcal{X}} x p_{X|A}(x).$$

If X is a *discrete* random variable, Y is a random variable defined as $Y = g(X)$, and A is an event, then the conditional expected value of Y is given by

$$\mathbb{E}[Y|A] = \mathbb{E}[g(X)|A] = \sum_{x \in \mathcal{X}} g(x) p_{X|A}(x).$$

If X and Y are two *discrete* random variables, W is defined as $W = g(X, Y)$, and A is an event, then the conditional expected value of W is given by

$$\mathbb{E}[W|A] = \mathbb{E}[g(X, Y)|A] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} g(x, y) p_{XY|A}(x, y).$$

For a *discrete* random variable X resulting from an experiment with partition $A_1, \ldots, A_m$, the expected value of X can be expressed as

$$\mathbb{E}[X] = \sum_{i=1}^{m} \mathbb{E}[X|A_i] P(A_i).$$

For a *continuous* random variable X and an event A with probability $P(A) > 0$, the **conditional expected value** of X is defined as

$$\mathbb{E}[X|A] = \int_{\mathcal{X}} x f_{X|A}(x) dx.$$

If X is a *continuous* random variable, Y is defined as $Y = g(X)$, and A is an event, then the conditional expected value of Y is given by

$$\mathbb{E}[Y|A] = \mathbb{E}[g(X)|A] = \int_{\mathcal{X}} g(x) f_{X|A}(x) dx.$$

If X and Y are two *continuous* random variables, W is defined as $W = g(X, Y)$, and A is an event, then the conditional expected value of W is given by

$$\mathbb{E}[W|A] = \mathbb{E}[g(X, Y)|A] = \int_{\mathcal{Y}} \int_{\mathcal{X}} g(x, y) f_{XY|A}(x, y) dx dy.$$

For a *continuous* random variable X resulting from an experiment with partition $A_1, \ldots, A_m$, the expected value of X can be expressed as

$$\mathbb{E}[X] = \sum_{i=1}^{m} \mathbb{E}[X|A_i] P(A_i).$$

7.1.5 Conditional Variance

For a random variable X (discrete or continuous) and an event A with probability $P(A) > 0$, the **conditional variance** of X is defined as

$$\text{Var}(X|A) = \mathbb{E}\left[\Big(X - \mathbb{E}[X|A]\Big)^2 \Big| A\right] = \mathbb{E}[X^2|A] - \Big(\mathbb{E}[X|A]\Big)^2.$$

The **conditional standard deviation** is defined in terms of the variance as

$$\sigma_{X|A} = \sqrt{\text{Var}(X|A)}.$$

7.2 Conditioning by a Random Variable

7.2.1 Conditional Probability Mass Function

For two discrete random variables X and Y, we can define two **conditional probability mass functions (PMFs)**, which are given by

$$p_{X|Y}(x|y) = P(\{X = x|Y = y\}) = \frac{p_{XY}(x,y)}{p_Y(y)}$$

and

$$p_{Y|X}(y|x) = P(\{Y = y|X = x\}) = \frac{p_{XY}(x,y)}{p_X(x)}.$$

The **conditional PMF theorem** says that the joint PMF can be expressed as

$$p_{XY}(x,y) = p_{X|Y}(x|y)p_Y(y) = p_{Y|X}(y|x)p_X(x).$$

If X and Y are independent random variables, then

$$p_{X|Y}(x|y) = p_X(x) \quad \text{and} \quad p_{Y|X}(y|x) = p_Y(y)$$

7.2.2 Conditional Cumulative Distrubtion Function

For two random variables X and Y, we can define two **conditional conditional cumulative distribution functions (CDFs)**, which are given by

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

and

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

7.2.3 Conditional Probability Density Function

For two continuous random variables X and Y, two **conditional probability density functions (PDFs)** are defined in terms of the conditional CDFs as

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

and

$$f_{Y|X}(y|x) = \frac{d}{dy}F_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)}.$$

The **conditional PDF theorem** says that the joint PDF can be expressed as

$$f_{XY}(x,y) = f_{X|Y}(x|y)f_Y(y) = f_{Y|X}(y|x)f_X(x).$$

If X and Y are independent random variables, then

$$f_{X|Y}(x|y) = f_X(x) \quad \text{and} \quad f_{Y|X}(y|x) = f_Y(y).$$

7.2.4 Conditional Expected Value

If X and Y are two *discrete* random variables, then the **conditional expected value** of X given that $Y = y$ is defined as

$$\mathbb{E}[X|Y=y] = \sum_{x\in\mathcal{X}} x p_{X|Y}(x|y).$$

If X and Y are two *discrete* random variables and W is defined as $W = g(X,Y)$, then the conditional expected value of W given $Y = y$ is defined as

$$\mathbb{E}[W|Y=y] = \mathbb{E}[g(X,Y)|Y=y] = \sum_{x\in\mathcal{X}} g(x,y)p_{X|Y}(x|y).$$

If X and Y are two *continuous* random variables, then the **conditional expected value** of X given taht $Y = y$ is defined as

$$\mathbb{E}[X|Y=y] = \int_{\mathcal{X}} x f_{X|Y}(x|y)dx.$$

If X and Y are two *continuous* random variables and W is defined as $W = g(X,Y)$, then the conditional expected value of W given $Y = y$ is defined as

$$\mathbb{E}[W|Y=y] = \mathbb{E}[g(X,Y)|Y=y] = \int_{\mathcal{X}} g(x,y)f_{X|Y}(x|y)dx.$$

If the random variables X and Y are independent, then

$$\mathbb{E}[X|Y=y] = E[X],\ \forall y\in\mathcal{Y} \quad \text{and} \quad \mathbb{E}[Y|X=x] = E[Y],\ \forall x\in\mathcal{X}.$$

7.2.5 Iterated Expectation & Tower Property

The conditional expected value $\mathbb{E}[X|Y]$ is a function of the random variable Y such that if $Y = y$, then $\mathbb{E}[X|Y] = \mathbb{E}[X|Y = y]$. Because $\mathbb{E}[X|Y]$ is a function of a random variable, it is also a random variable. To calculate the expected value of $g(X)$, we can compute the **iterated expected value**:

$$\mathbb{E}[g(X)] = \mathbb{E}\big[\mathbb{E}[g(X)|Y]\big].$$

If X is a discrete random variable, then this expected value is given by

$$\mathbb{E}[g(X)] = \sum_{y\in\mathcal{Y}} \mathbb{E}\big[g(X)|Y = y\big]p_Y(y).$$

If X is a continuous random variable, then this expected value is given by

$$\mathbb{E}[g(X)] = \int_{\mathcal{Y}} \mathbb{E}\big[g(X)|Y = y\big]f_Y(y)dy.$$

Similar to the iterated expected value, we also have the **tower property**, which says that for any two random variables X and Y and any function f,

$$\mathbb{E}[f(Y)X] = \mathbb{E}\big[f(Y)\mathbb{E}[X|Y]\big].$$

7.2.6 Conditional Variance

If X and Y are two random variables, the variance of X given that $Y = y$ is

$$\begin{aligned}\mathrm{Var}(X|Y = y) &= \mathbb{E}\left[\Big(X - \mathbb{E}[X|Y = y]\Big)^2\Big|Y = y\right] \\ &= \mathbb{E}[X^2|Y = y] - \big(\mathbb{E}[X|Y = y]\big)^2.\end{aligned}$$

The conditional variance $\mathrm{Var}(X|Y)$ is a function of the random variable Y such that if $Y = y$, then $\mathrm{Var}(X|Y) = \mathrm{Var}(X|Y = y)$. Because $\mathrm{Var}(X|Y)$ is a function of a random variable, it is also a random variable. Using iterated expectation, we can write the **minimum mean-square error (MMSE)** of X given Y:

$$\mathbb{E}[\mathrm{Var}(X|Y)] = \mathbb{E}\left[\Big(X - \mathbb{E}[X|Y]\Big)^2\right].$$

The **law of total variance** says that the variance of X can be expressed as

$$\mathrm{Var}(X) = \mathbb{E}\big[\mathrm{Var}(X|Y)\big] + \mathrm{Var}\big(\mathbb{E}[X|Y]\big).$$

Rearranging this equation, we see that we can express the **minimum mean-square error (MMSE)** of X given Y as

$$\mathbb{E}[\mathrm{Var}(X|Y)] = \mathrm{Var}(X) - \mathrm{Var}\big(\mathbb{E}[X|Y]\big).$$

Part IV

Extensions of Random Variables

Chapter 8

Random Vectors

8.1 Random Vector

A **random vector** with n dimensions is a concise representation of a set of n random variables. Often we express the n-dimensional random vector $\boldsymbol{X}$ as

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

where each element X_i is a random variable. A realization or sample value of the random vector $\boldsymbol{X}$ is another vector $\boldsymbol{x}$, which can be expressed as

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix},$$

where the ith element x_i is a sample value of the random variable X_i.

8.2 Probability Distributions

8.2.1 Probability Mass Functions

The probability mass function (PMF) for a discrete random vector $\boldsymbol{X}$ is

$$p_{\mathbf{X}}(\boldsymbol{x}) = p_{X_1 \ldots X_n}(x_1, \ldots, x_n).$$

If $\boldsymbol{X}$ is a discrete random vector and $\boldsymbol{W} = g(\boldsymbol{X})$, the PMF of $\boldsymbol{W}$ is given by

$$p_{\mathbf{W}}(\boldsymbol{w}) = \sum_{\boldsymbol{x}: g(\boldsymbol{x}) = \boldsymbol{w}} p_{\mathbf{X}}(\boldsymbol{x}).$$

The joint PMF for a pair of discrete random vectors, $\boldsymbol{X}$ and $\boldsymbol{Y}$, is defined as

$$p_{\mathbf{X},\mathbf{Y}}(\boldsymbol{x},\boldsymbol{y}) = p_{X_1 \ldots X_n Y_1 \ldots Y_m}(x_1,\ldots,x_n,y_1,\ldots,y_m).$$

The random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ are said to be independent if and only if

$$p_{\mathbf{X},\mathbf{Y}}(\boldsymbol{x},\boldsymbol{y}) = p_{\mathbf{X}}(\boldsymbol{x}) p_{\mathbf{Y}}(\boldsymbol{y}).$$

8.2.2 Cumulative Distribution Functions

The cumulative distribution function (CDF) for a random vector $\boldsymbol{X}$ is

$$F_{\mathbf{X}}(\boldsymbol{x}) = F_{X_1 \ldots X_n}(x_1,\ldots,x_n).$$

If $\boldsymbol{X}$ is a discrete random vector and $\boldsymbol{W} = g(\boldsymbol{X})$, the CDF of W is given by

$$F_{\mathbf{W}}(\boldsymbol{w}) = \sum_{\boldsymbol{x}:g(\boldsymbol{x})\leq\boldsymbol{w}} p_{\mathbf{X}}(\boldsymbol{x}).$$

If $\boldsymbol{X}$ is a continuous random vector and $\boldsymbol{W} = g(\boldsymbol{X})$, the CDF of $\boldsymbol{W}$ is given by

$$F_{\mathbf{W}}(\boldsymbol{w}) = \int_{g(\boldsymbol{x})\leq\boldsymbol{w}} f_{\mathbf{X}}(\boldsymbol{x}) d\boldsymbol{x}.$$

The joint CDF for a pair of random vectors, $\boldsymbol{X}$ and $\boldsymbol{Y}$, is defined as

$$F_{\mathbf{X},\mathbf{Y}}(\boldsymbol{x},\boldsymbol{y}) = F_{X_1 \ldots X_n Y_1 \ldots Y_m}(x_1,\ldots,x_n,y_1,\ldots,y_m).$$

8.2.3 Probability Density Functions

The probability density function (PDF) for a continuous random vector $\boldsymbol{X}$ is

$$f_{\mathbf{X}}(\boldsymbol{x}) = f_{X_1 \ldots X_n}(x_1,\ldots,x_n).$$

If $\boldsymbol{X}$ is a continuous random vector and the random variable $\boldsymbol{W} = g(\boldsymbol{X})$ has the CDF $F_{\mathbf{W}}$, the PDF of $\boldsymbol{W}$ can be determined from the CDF of $\boldsymbol{W}$:

$$f_{\mathbf{W}}(\boldsymbol{w}) = \nabla_{\mathbf{w}} F_{\mathbf{W}}(\boldsymbol{w}).$$

The joint PDF for a pair of continuous random vectors, $\boldsymbol{X}$ and $\boldsymbol{Y}$, is defined as

$$f_{\mathbf{X},\mathbf{Y}}(\boldsymbol{x},\boldsymbol{y}) = f_{X_1 \ldots X_n Y_1 \ldots Y_m}(x_1,\ldots,x_n,y_1,\ldots,y_m).$$

The random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ are said to be independent if and only if

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

8.3 Properties of Random Vectors

8.3.1 Expected Value Vector

The expected value of a random vector is a column vector whose elements are the expected values of the individual random variables, which we express as

$$\boldsymbol{\mu}_{\mathbf{X}} = \mathbb{E}[\boldsymbol{X}] = \begin{bmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_n] \end{bmatrix}.$$

If $\boldsymbol{X}$ is a discrete random vector, its expected value is determined by its PMF:

$$\mathbb{E}[\boldsymbol{X}] = \sum_{x_1 \in \mathcal{X}_1} \cdots \sum_{x_n \in \mathcal{X}_n} \boldsymbol{x} p_{\mathbf{X}}(\boldsymbol{x}).$$

If $\boldsymbol{X}$ is a continuous random vector, its expected value is defined by its PDF:

$$\mathbb{E}[\boldsymbol{X}] = \int_{\mathcal{X}_1} \cdots \int_{\mathcal{X}_n} \boldsymbol{x} f_{\mathbf{X}}(\boldsymbol{x}) dx_1 \dots dx_n.$$

If $\boldsymbol{X}$ is a discrete random vector and $\boldsymbol{Y} = g(\boldsymbol{X})$, then $\boldsymbol{Y}$ has the expected value

$$\mathbb{E}[\boldsymbol{Y}] = \mathbb{E}[g(\boldsymbol{X})] = \sum_{x_1 \in \mathcal{X}_1} \cdots \sum_{x_n \in \mathcal{X}_n} g(\boldsymbol{x}) p_{\mathbf{X}}(\boldsymbol{x}).$$

If $\boldsymbol{X}$ is a continuous random vector and $\boldsymbol{Y} = g(\boldsymbol{X})$, $\boldsymbol{Y}$ has the expected value

$$\mathbb{E}[\boldsymbol{Y}] = \mathbb{E}[g(\boldsymbol{X})] = \int_{\mathcal{X}_1} \cdots \int_{\mathcal{X}_n} g(\boldsymbol{x}) f_{\mathbf{X}}(\boldsymbol{x}) dx_1 \dots dx_n.$$

When the components of $\boldsymbol{X}$ are all independent, the following property holds:

$$\mathbb{E}[g_1(X_1) \dots g_n(X_n)] = \prod_{i=1}^{n} \mathbb{E}[g_i(X_i)].$$

8.3.2 Covariance & Correlation Matrices

The covariance of an n-dimensional random vector $\boldsymbol{X}$ is an $n \times n$ matrix $\boldsymbol{\Sigma}_{\mathbf{X}}$ whose (i,j)th element is $\text{Cov}(X_i, X_j)$. In vector notation, we write this as

$$\boldsymbol{\Sigma}_{\mathbf{X}} = \mathbb{E}\left[\left(\boldsymbol{X} - \boldsymbol{\mu}_{\mathbf{X}} \right) \left(\boldsymbol{X} - \boldsymbol{\mu}_{\mathbf{X}} \right)^T \right].$$

From this definition of the covariance, we can see that the **covariance matrix** $\boldsymbol{\Sigma}_{\mathbf{X}}$ is always positive semidefinite (PSD), meaning $\boldsymbol{v}^T \boldsymbol{\Sigma}_{\mathbf{X}} \boldsymbol{v} \geq 0$ for all $\boldsymbol{v} \in \mathbb{R}^n$.

The correlation of an n-dimensional random vector X is an $n \times n$ matrix $\boldsymbol{R}_{\boldsymbol{X}}$ whose (i,j)th element is $\mathbb{E}[X_i X_j]$. In vector notation, we write this as

$$\boldsymbol{R}_{\boldsymbol{X}} = \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^T].$$

For a random vector $\boldsymbol{X}$ with **correlation matrix** $\boldsymbol{R}_{\boldsymbol{X}}$, covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{X}}$, and expected value vector $\boldsymbol{\mu}_{\boldsymbol{X}}$, we can write the following relationship:

$$\boldsymbol{R}_{\boldsymbol{X}} = \boldsymbol{\Sigma}_{\boldsymbol{X}} + \boldsymbol{\mu}_{\boldsymbol{X}} \boldsymbol{\mu}_{\boldsymbol{X}}^T.$$

8.3.3 Linear Transformation

Consider the n-dimensional random vector $\boldsymbol{X}$ with correlation matrix $\boldsymbol{R_X}$, covariance matrix $\boldsymbol{\Sigma_X}$, and expected value vector $\boldsymbol{\mu_X}$. If $\boldsymbol{A}$ is an $m \times n$ matrix and $\boldsymbol{b}$ is an m-dimensional vector, then the random vector $\boldsymbol{Y} = \boldsymbol{Ax} + \boldsymbol{b}$ has the following mean, correlation matrix, and covariance matrix:

$$\boldsymbol{\mu_Y} = \boldsymbol{A\mu_X} + \boldsymbol{b}$$

$$\boldsymbol{R_Y} = \boldsymbol{AR_X A}^T + (\boldsymbol{A\mu_X})\boldsymbol{b}^T + \boldsymbol{b}(\boldsymbol{A\mu_X})^T + \boldsymbol{bb}^T$$

$$\boldsymbol{\Sigma_Y} = \boldsymbol{A\Sigma_X A}^T$$

8.4 Gaussian Random Vectors

The random vector $\boldsymbol{X}$ is a Gaussian random vector with expected value vector $\boldsymbol{\mu_X}$ and covariance matrix $\boldsymbol{\Sigma_X}$ if and only if its PDF can be expressed as

$$f_{\mathbf{X}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}\left|\boldsymbol{\Sigma_X}\right|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu_X})^T \boldsymbol{\Sigma_X}^{-1}(\boldsymbol{x} - \boldsymbol{\mu_X})\right).$$

The Gaussian random vector $\boldsymbol{X}$ has independent components if and only if all of its components are uncorrelated, meaning $\mathrm{Cov}(X_i, X_j) = 0$ for all $i \neq j \in [1, n]$. Therefore, a Gaussian random vector $\boldsymbol{X}$ has independent components if and only if $\boldsymbol{\Sigma_X}$ is a diagonal matrix, whose diagonal elements are $\sigma_i^2 = \mathrm{Var}(X_i)$, i.e.

$$\boldsymbol{\Sigma_X} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2) = \begin{bmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_n^2 \end{bmatrix}.$$

Consider an n-dimensional Gaussian random variable $\boldsymbol{X}$ with expected value μ_X and covariance $\boldsymbol{\Sigma_X}$. If $\boldsymbol{A}$ is an $m \times n$ matrix and $\boldsymbol{b}$ is an m-dimensional vector, then the random vector $\boldsymbol{Y} = \boldsymbol{AX} + \boldsymbol{b}$ is an m-dimensional Gaussian random vector with mean vector and covariance matrix given by

$$\boldsymbol{\mu_Y} = \boldsymbol{A\mu_X} + \boldsymbol{b} \quad \text{and} \quad \boldsymbol{\Sigma_Y} = \boldsymbol{A\Sigma_X A}^T.$$

8.4.1 Standard Normal Random Vector

The n-dimensional standard normal random vector $\boldsymbol{Z}$ is the n-dimensional Gaussian random vector with expected value vector $\boldsymbol{\mu_Z} = \boldsymbol{0_n}$ and covariance matrix $\boldsymbol{\Sigma_Z} = \boldsymbol{I_n}$. This means each component of $\boldsymbol{Z}$ is a random variable Z_i with an expected value of zero and variance of one. Furthermore, the covariance between all of the random variables is zero, which implies that all of the random variables are uncorrelated, and thus independent.

We can easily transform between general Gaussian random vectors and standard normal random vectors. Assume $\boldsymbol{X}$ is an n-dimensional Gaussian random vector

with expected value $\boldsymbol{\mu_X}$ and covariance $\boldsymbol{\Sigma_X}$. Because the covariance matrix is postive semidefinite, we can define an $n \times n$ matrix $\boldsymbol{A}$ such that $\boldsymbol{AA}^T = \boldsymbol{\Sigma_X}$. The random vector $\boldsymbol{Z} = \boldsymbol{A}^{-1}(\boldsymbol{X} - \boldsymbol{\mu_X})$ is an n-dimensional standard normal random vector. Similarly, given the n-dimensional standard normal random vector $\boldsymbol{Z}$, an invertible $n \times n$ matrix $\boldsymbol{A}$ and an n-dimensional vector $\boldsymbol{b}$, the random vector $\boldsymbol{X} = \boldsymbol{AZ} + \boldsymbol{b}$ is an n-dimensional Gaussian random vector with expected value $\boldsymbol{\mu_X} = \boldsymbol{b}$ and covariance $\boldsymbol{\Sigma_X} = \boldsymbol{AA}^T$.

Chapter 9

Concentration Inequalities

9.1 Sample Mean

Let $X_1, \ldots, X_n$ be independent and identically distributed random samples drawn from a population with variance σ^2 and expected value μ. The **sample mean** of this set of n samples is denoted $M_n(X)$ and is defined as

$$M_n(X) = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Note that the sample mean is a random variable because it is a function of random variables. Its expected value and variance can be expressed as

$$\mathbb{E}[M_n(X)] = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} X_i\right] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[X_i] = \frac{1}{n} \sum_{i=1}^{n} \mu = \frac{1}{n}(n)\mu = \mu$$

and

$$\mathrm{Var}(M_n(X)) = \mathrm{Var}\left(\frac{1}{n} \sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}(X_i) = \frac{1}{n^2} \sum_{i=1}^{n} \sigma^2 = \frac{1}{n}\sigma^2.$$

The expected value of the sample mean is simply the mean of the population, and the variance of the sample mean is the variance of the population divided by the number of samples. As the number of samples approaches infinity, the variance of the sample mean approaches zero. Therefore, the sample mean converges to the population mean as the number of samples approaches infinity.

Now suppose that $X_1, \ldots, X_n$ are correlated random samples drawn from a population with variance σ^2, expected value μ, and correlation coefficient $\rho > 0$. For this case, the expected value of the sample mean would not change. However,

the variance for the sample mean would now be given by

$$\begin{aligned}
\text{Var}(M_n(X)) &= \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\text{Var}\left(\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\text{Cov}\left(\sum_{i=1}^{n} X_i,\ \sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n}\text{Var}(X_i) + \sum_{i=1}^{n}\sum_{j\neq i}\text{Cov}(X_i, X_j)\right) \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n}\sigma^2 + \sum_{i=1}^{n}\sum_{j\neq i}\rho\sigma^2\right) \\
&= \frac{1}{n^2}\left(n\sigma^2 + n(n-1)\rho\sigma^2\right) \\
&= \frac{1+(n-1)\rho}{n}\sigma^2.
\end{aligned}$$

Notice that the variance of the sample mean no longer approaches zero as the number of samples approaches infinity. Therefore, the sample mean may not converge to the population mean as the number of samples approaches infinity.

9.2 Probability Bounds

9.2.1 Markov Inequality

Theorem: The **Markov inequality** says that for a non-negative random variable X and a positive constant $c > 0$, the following inequality holds:

$$P(\{X \geq c\}) \leq \frac{\mathbb{E}[X]}{c}.$$

Proof: From the definition of the indicator function, we can write:

$$\mathbb{1}\{X \geq c\} = \begin{cases} 1 & \text{if } X \geq c \\ 0 & \text{if } X < c \end{cases}$$

$$c \cdot \mathbb{1}\{X \geq c\} = \begin{cases} c & \text{if } X \geq c \\ 0 & \text{if } X < c \end{cases}$$

Now notice that because X is a non-negative random variable,

$$c \cdot \mathbb{1}\{X \geq c\} \leq X.$$

Because the expected value function is monotonically increasing,

$$\mathbb{E}[c \cdot \mathbb{1}\{X \geq c\}] \leq \mathbb{E}[X]$$

From the linearity of expectation, we can see that

$$c \cdot \mathbb{E}[\mathbb{1}\{X \geq c\}] \leq \mathbb{E}[X].$$

Because c is strictly positive, we can divide both sides by c, so

$$\mathbb{E}[\mathbb{1}\{X \geq c\}] \leq \frac{\mathbb{E}[X]}{c}.$$

Finally, from our discussion of common discrete distributions, we know the expected value of an indicator function, which leads us to the desired result:

$$P(\{X \geq c\}) \leq \frac{\mathbb{E}[X]}{c}.$$

Extension: Note that there are several other inequalities that can be derived from the Markov inequality. One useful extension says that for a random variable X with finite variance and $\mathbb{E}[X] = 0$,

$$P(\{X \geq c\}) \leq \frac{\mathbb{E}[X^2]}{\mathbb{E}[X^2] + c^2}, \ \forall c \geq 0.$$

9.2.2 Chebyshev Inequality

Theorem: The **Chebyshev inequality** says that for any arbitrary random variable X with expected value μ_X and variance $\text{Var}(X)$ and for an arbitrary positive constant $c > 0$, the following inequality holds:

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

Proof: This inequality comes directly from the Markov inequality.

9.2.3 Chernoff Bound

Theorem: The **Chernoff bound** says that that for any arbitrary random variable X with moment generating function $\phi_X(s)$ and any constant c,

$$P(\{X \geq c\}) \leq \min_{s \geq 0} e^{-sc} \phi_X(s).$$

Proof: This inequality comes directly from the Markov inequality.

It's interesting to consider the implications of the Chernoff bound for Gaussian random variables. For the Gaussian random variable $X \sim N(\mu, \sigma^2)$, it says

$$\begin{aligned} P(\{X \geq c\}) &\leq \min_{s \geq 0} \left\{ e^{-sc} \cdot \exp\Big(\frac{s\mu + s^2\sigma^2}{2} \Big) \right\} \\ &= \min_{s \geq 0} \left\{ \exp\Big(\frac{s(\mu - 2c) + s^2\sigma^2}{2} \Big) \right\}. \end{aligned}$$

To find the optimal value, $\hat{s}$, that minimizes this objective, we can take the derivative of the objective with respect to s, plug in $\hat{s}$, and set the expression equal to zero. Following this approach, we get the following result:

$$\exp\Big(\frac{\hat{s}(\mu-2c)+\hat{s}^2\sigma^2}{2}\Big)\Big(\frac{\mu}{2}-c+\sigma^2\hat{s}\Big)=0$$

$$\Big(\frac{\mu}{2}-c+\sigma^2\hat{s}\Big)=0$$

$$\hat{s}=\frac{2c-\mu}{2\sigma^2}$$

Plugging this expression for $\hat{s}$ into our previous inequality, we get

$$\begin{aligned}
P(\{X\geq c\}) &\leq \min_{s\geq 0}\left\{\exp\Big(\frac{s(\mu-2c)+s^2\sigma^2}{2}\Big)\right\} \\
&= \exp\Big(\frac{\hat{s}(\mu-2c)+\hat{s}^2\sigma^2}{2}\Big) \\
&= \exp\Big(\frac{-\frac{(\mu-2c)^2}{2\sigma^2}+\frac{(\mu-2c)^2}{4\sigma^2}}{2}\Big) \\
&= \exp\Big(-\frac{(\mu-2c)^2}{8\sigma^2}\Big).
\end{aligned}$$

Because the Gaussian random variable X is symmetric, we can also write

$$P(\{|X|\geq c\})\leq 2\exp\Big(-\frac{(\mu-2c)^2}{8\sigma^2}\Big).$$

9.3 Law of Large Numbers

9.3.1 Convergence of Random Variables

There are three commonly used modes of convergence for random variables:

1. **Almost Sure Convergence**

   For a sequence of random variables $X_1, X_2, \ldots$, we say that X_n converges to the random variable X almost surely if

   $$P\left(\left\{\lim_{n\to\infty} X_n = X\right\}\right)=1.$$

   We can equivalently write this expression as

   $$P\left(\left\{\omega\in\Omega:\lim_{n\to\infty} X_n(\omega)=X(\omega)\right\}\right)=1.$$

   This means that the values of X_n approach the value of X, in the sense that events for which X_n does not converge to X have probability zero.

2. **Convergence in Probability**

   For a sequence of random variables $X_1, X_2, \ldots$, we say that X_n converges to the random variable X in probability if for every constant $\epsilon > 0$,

$$\lim_{n\to\infty} P\left(\{|X_n - X| > \epsilon\}\right) = 0.$$

3. **Convergence in Distribution**

   For a sequence of random variables $X_1, X_2, \ldots$ with corresponding CDFs $F_{X_1}, F_{X_2}, \ldots$, we say that X_n converges to the random variable X with CDF F_X in distribution if for every $x \in \mathbb{R}$ where the CDFs are continuous,

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x).$$

Note that almost sure convergence implies convergence in probability, and convergence in probability implies convergence in distribution. These implications do not hold in the opposite direction.

9.3.2 Weak Law of Large Numbers

Suppose that the random variable X represents a population from which samples are drawn. Let μ_X be the expected value of X, σ^2 be the variance, and $M_n(X)$ be the sample mean. The **weak law of large numbers** for a finite number of samples says that for any positive constant $c > 0$, the following holds:

$$(1)\ P\left(\{|M_n(X) - \mu_X| \geq c\}\right) \leq \frac{\sigma^2}{nc^2}$$

and

$$(2)\ P\left(\{|M_n(X) - \mu_X| < c\}\right) \geq 1 - \frac{\sigma^2}{nc^2}.$$

These inequalities come directly from the Chebyshev inequality. From our definitions of convergence, the weak law of large numbers says that the empirical mean $M_n(X)$ converges to the true mean μ_X in probability.

Taking the limit of these inequalities as n goes to infinity, we get the weak law of large numbers for an infinite number of samples:

$$\lim_{n\to\infty} P\left(\{|M_n(X) - \mu_X| \geq c\}\right) = 0$$

$$\lim_{n\to\infty} P\left(\{|M_n(X) - \mu_X| < c\}\right) = 1$$

This finding aligns with our expectations. If $M_n(X)$ is the average of a very large sample of independent and identically distributed random variables, then we expect that any given realization of $M_n(X)$ is very close to the expected value μ_X. As the number of samples goes to infinity, we expect that any realization from $M_n(X)$ will be equal to the expected value of the individual random variables. Therefore, as the number of samples goes to infinity, $P(\{|M_n(X) - \mu_X| > 0\}) = 0$.

9.3.3 Strong Law of Large Numbers

The **strong law of large numbers** says that for any random variable X with expected value μ_X and sample mean $M_n(X)$, the sample mean converges to the true mean μ_X almost surely. Note that the strong law of large numbers implies the weak law because almost sure convergence implies convergence in probability. This law does not say that anomalies cannot happen for a very large number n, but with probability one, these anomalies will not happen.

9.4 Central Limit Theorem

Let $X_1, \ldots, X_n$ be independent and identically distributed random variables with variance $\text{Var}(X_i) = \sigma^2 < \infty$ and expected value $\mathbb{E}[X_i] = \mu$. Now consider a new random variable S_n, which is defined as the following sum:

$$S_n = \frac{1}{\sqrt{n\sigma^2}} \left(\sum_{i=1}^{n} X_i - \mu \right).$$

The **central limit theorem** says that S_n converges to the standard normal random variable $Z \sim \mathcal{N}(0, 1)$ in distribution, meaning that for all $x \in \mathbb{R}$,

$$P(\{S_n \leq x\}) \to \phi(x).$$

The central limit theorem is useful because it says that when independent random variables are summed up, their properly normalized sum tends toward a standard normal distribution, even if the original variables themselves are not normally distributed. This implies that probabilistic and statistical methods that work for normal distributions can be applied to many problems involving other types of distributions, assuming a large enough sample size.