EECS126: Probability Theory and Random Processes

UC Berkeley

Kelvin Lee

December 6, 2020

These are course notes for UC Berkeley's Spring 2020 EECS126 Probability Theory and Random Processes, instructed by Professor Kannan Ramchandran.

Contents

1 Sample Space and Probability **3**
- 1.1 Probabilistic Models . . . 3
- 1.2 Probability Space . . . 3
  - 1.2.1 Properties of Probability Laws . . . 4
- 1.3 Discrete Uniform Probability Space . . . 4
  - 1.3.1 Birthday Paradox . . . 5
- 1.4 Conditional Probability . . . 5
- 1.5 Independence . . . 5
  - 1.5.1 Conditional Independence . . . 5
- 1.6 Law of Total Probability . . . 6
- 1.7 Bayes' Rule . . . 6

2 Discrete Random variables **7**
- 2.1 Expectation . . . 8
  - 2.1.1 Linearity of Expectation . . . 8
- 2.2 Variance . . . 8
  - 2.2.1 Covariance . . . 10
  - 2.2.2 Correlation . . . 10
- 2.3 Discrete Probability Distribution . . . 10
  - 2.3.1 Bernoulli Distribution . . . 10
  - 2.3.2 Binomial Distribution . . . 11
  - 2.3.3 Hypergeometric Distribution . . . 11
  - 2.3.4 Geometric Distribution . . . 11
  - 2.3.5 Poisson Distribution . . . 12
- 2.4 Conditioning of Random Variables . . . 14
  - 2.4.1 Conditional Expectation . . . 15

3 Continuous Probability **17**
- 3.1 Continuous Random Variables . . . 17
- 3.2 Expectation and Variance . . . 17
- 3.3 Continuous Probability Distribution . . . 17
  - 3.3.1 Exponential Random Variable . . . 17
  - 3.3.2 Cumulative Distribution Functions . . . 18
- 3.4 Normal Random Variables . . . 19
- 3.5 Joint PDFs of Multiple Random Variables . . . 20

3.6 Joint CDFs . . . 20
3.7 Conditioning . . . 20
3.7.1 Continuous Bayes' Rule . . . 21

4 More on Random Variables 22
4.1 Derived Distributions . . . 22
4.2 Convolution . . . 23
4.3 Law of Iterated Expectations . . . 23
4.4 Law of Total Variance . . . 24
4.5 Order Statistics . . . 25

5 Moment Generating Functions (Transforms) 26
5.1 Inversions of transforms . . . 27
5.2 Sums of Independent Random Variables . . . 28

6 Limit Theorems 30
6.1 Markov's Inequality . . . 30
6.2 Chebyshev's Inequality . . . 30
6.3 Chernoff Bounds . . . 30
6.4 Weak Law of Large Numbers . . . 31
6.5 Convergence in Probability . . . 32
6.6 The Central Limit Theorem . . . 34

7 Information Theory 36

1 Sample Space and Probability

1.1 Probabilistic Models

A **probabilistic model** is a mathematical description of an uncertain situation. The elements of a probabilistic model includes

- **sample space** Ω : set of all all possible outcomes of an experiment.
- **probability law:** assigns to a set A of possible outcomes (**event**) a nonnegative value $\mathbf{P}(A)$ (probability of A) that encodes the knowledge about the likelihood of the elements of A.

A recap of all basic terminologies:

Definition 1 (Experiment)**.** An **experiment** is a procedure that yields one of a given set of possible outcomes.

Definition 2 (Sample space)**.** The **sample space** of the experiment is the set of possible outcomes.

Definition 3 (Sample point)**.** A **sample point** is an element of the sample space.

Definition 4 (Event)**.** An **event** is a subset of the sample space.

1.2 Probability Space

Definition 5 (Probability Space)**.** The **probability space** is defined by the triple $(\Omega, \mathcal{F}, \mathbf{P})$ where Ω is the *sample space*, $\mathcal{F} \subseteq \Omega$ is the *event space* and $\mathbf{P}$ is the *probability function*, satisfying the following axioms:

Probability Axioms (Kolmogorov):

- **Nonnegativity:** for all sample points $\omega \in \Omega$,

$$\mathbf{P}(\omega) \geq 0.$$

- **Additivity:** any countable sequence of **disjoint sets** (mutually exclusive events) $E_1, E_2, \ldots$ satisfies

$$\mathbf{P}\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mathbf{P}\left(E_i\right).$$

- **Normalization:** the sum of all probabilities must be 1, thus

$$\sum_{\omega \in \Omega} \mathbf{P}(\omega) = \mathbf{P}(\Omega) = 1.$$

Definition 6 (Probability)**.** For any event $A \subseteq \Omega$, we define the **probability** of A to be

$$\mathbf{P}(A) = \sum_{\omega \in A} \mathbf{P}(\omega).$$

1.2.1 Properties of Probability Laws

- $\mathbf{P}(\emptyset) = 0$.
- $\mathbf{P}(\overline{A}) = 1 - \mathbf{P}(A)$, where $\overline{A}$ (or A^c) is the **complement** of A.
- $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(A) - \mathbf{P}(A \cap B)$.
- If $A \subseteq B$, then $\mathbf{P}(A) \leq \mathbf{P}(B)$.

1.3 Discrete Uniform Probability Space

Theorem 7 (Discrete Uniform Probability Law)**.** In a uniform probability space, all sample points have the same probability $\frac{1}{|\Omega|}$. Thus the probability of an event A is

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

Remark. For uniform spaces, computing probabilities is simply counting sample points.

Example 1.1 (Poker Hands)**.** Consider shuffling a deck of cards and dealing a poker hand. In this case, the sample space $\Omega = \{\text{all possible poker hands}\}$. Hence, $|\Omega| = \binom{52}{5}$. Assuming that the probability of each outcome is equally likely and so we have an uniform probability space.
Let A be the event that the poker hand is a flush (same suit). Since the probability space is uniform, computing $\mathbf{P}(A)$ reduces to simply computing $|A|$, the number of poker hands that are flushes. There are 13 cards in each suit, so the number of flushes in each suit is $\binom{13}{5}$. The total number of flushes is therefore $4 \cdot \binom{13}{5}$. Then we have

$$\mathbf{P}\left(\text{hand is a flush}\right) \approx 0.002.$$

Example 1.2 (Balls and Bins)**.** Consider the experiment of throwing 20 labelled balls into 10 labeled bins. Assume that each ball is equally likely to land in any bin.
The sample space Ω is equal to $\{(b_1, b_2, \ldots, b_{20}) : 1 \leq b_i \leq 10 \text{ for each } i = 1, \ldots, 20\}$, where the component b_i denotes the bin in which ball i lands. Then $|\Omega| = 10^{20}$, since each element b_i in the sequence has 10 possible choices and there are 20 elements in the sequence. In general, throwing m balls into n bins gives a sample space of size n^m.

Let A be the event that bin 1 is empty. Since the probability space is uniform, we simply need to count how many outcomes have this property. This is exactly the number of ways all 20 balls can fall into the remaining nine bins, which is 9^{20}. Hence, $\mathbf{P}(A) = \frac{9^{20}}{10^{20}} = \left(\frac{9}{10}\right)^{20} \approx 0.12$. Let B be the event that bin 1 contains at least one ball. This event is the complement $\overline{A}$ of A. So $\mathbf{P}(B) = 1 - \mathbf{P}(A) \approx 0.88$. More generally, if we throw m balls into n bins, we have:

$$\mathbf{P}(\text{bin 1 is empty}) = \left(\frac{n-1}{n}\right)^m = \left(1 - \frac{1}{n}\right)^m.$$

1.3.1 Birthday Paradox

The **birthday paradox** examines the chances that two people in a group have the same birthday. It is called a "paradox" because it is counter-intuitive. Suppose there are 365 days in a year. Then $S = \{1, \ldots, 365\}$, and the experiment consists of drawing a sample of n elements from S, where the elements are the birth dates of n people in a group. Then $|\Omega| = 365^n$ because there are 365 possible birth dates for each person. Let A be the event that at least a pair of people have the same birthday. If we want to determine $\mathbf{P}(A)$, it is simpler to first compute the probability of the complement of A; i.e., $\mathbf{P}(\overline{A})$, where $\overline{A}$ is the event that no two people have the same birthday.

Since the probability space is uniform, we just need to determine $|\overline{A}|$, the number of ways for no two people to have the same birthday. There are 365 choices for the first person, 364 for the second, $\ldots, 365-n+1$ choices for the n-th person, for a total of $365 \times 364 \times \cdots \times (365-n+1)$ by the First Rule of Counting from previous section; we are sampling without replacement and the order matters. Thus we have

$$\mathbf{P}(\overline{A}) = \frac{|\overline{A}|}{|\Omega|} = \frac{365 \times 364 \times \cdots \times (365-n+1)}{365^n},$$

so $\mathbf{P}(A) = 1 - \mathbf{P}(\overline{A}) = 1 - \frac{365 \times 364 \times \cdots \times (365-n+1)}{365^n}$. Here $\mathbf{P}(A)$ is a function of n. As n increases $\mathbf{P}(A)$ increases. For example, with $n = 23$ people, you should be willing to bet that at least a pair of people have the same birthday, since $\mathbf{P}(A)$ is larger than 50%. For $n = 60$ people, $\mathbf{P}(A)$ is over 99%!

1.4 Conditional Probability

Conditional probability provides a way to reason about the outcome of an experiment, based on partial information. We wish to quantify the likelihood that the outcome also belongs to some other given event A by constructing a new probability law to take into account the available knowledge.

Definition 8 (Conditional Probability)**.** Let B be an event such that $\mathbf{P}(B) > 0$. The **conditional probability** of A given B, denoted by $\mathbf{P}(A|B)$ is defined as

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

1.5 Independence

Definition 9 (Independence)**.** Event A and B are **independent** if and only if $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$ or $\mathbf{P}(A|B) = \mathbf{P}(A)$.

Definition 10 (Independence of Several Events)**.** Events $\{A\}_{i=1}^n$ are **independent** if

$$\mathbf{P}\left(\bigcap_{i \in S} A_i\right) = \prod_{i \in S} \mathbf{P}(A_i).$$

1.5.1 Conditional Independence

Definition 11 (Conditional Independence)**.** Given event C such that $\mathbf{P}(C) > 0$, events A and B are called **conditionally independent** if

$$\mathbf{P}(A \cap B|C) = \mathbf{P}(A|C)\mathbf{P}(B|C),$$

or

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

1.6 Law of Total Probability

Theorem 12 (Law of Total Probability)**.**

$$\mathbf{P}(A) = \sum_n \mathbf{P}\left(A \cap B_n\right) = \sum_n \mathbf{P}\left(A \mid B_n\right) \mathbf{P}\left(B_n\right).$$

1.7 Bayes' Rule

Theorem 13 (Bayes' Rule)**.** Let $\{A\}_{i=1}^n$ be disjoint events that form a partition of the sample space, and that $\mathbf{P}(A_i) > 0$ for all i. Then, for any event B such that $\mathbf{P}(B) > 0$,

$$\mathbf{P}(A_i \cap B) = \frac{\mathbf{P}(A_i)\mathbf{P}(B|A_i)}{\mathbf{P}(B)} = \frac{\mathbf{P}(A_i)\mathbf{P}(B|A_i)}{\sum_i^n \mathbf{P}(A_i)\mathbf{P}(B|A_i)}.$$

2 Discrete Random variables

Definition 14 (Random variable)**.** A **random variable** X on a sample space Ω is a function $X : \Omega \to \mathbb{R}$ that assigns to each sample point $\omega \in \Omega$ a real number $X(\omega)$.

Remark (Functions of R.V.s are also R.V.s)**.** Let $Y = g(X)$. Then

$$\mathbb{P}(Y = y) = \sum_{x|g(x)=y} \mathbb{P}(X = x).$$

An R.V. itself is a function, and we know that the function of a function is also a function.

Definition 15. The **distribution** of a discrete random variable X is the collection of values $\{(x, \mathbb{P}(X = x)) : x \in \mathcal{X}\}$, where $\mathcal{X}$ is the set of all possible values taken by X.

Definition 16 (probability Mass Function)**.** The **probability mass function**, or PMF, of a discrete random variable X is a function mapping X's values to their associated probabilities. It is the function $p : \mathbb{R} \to [0, 1]$ defined by

$$p_X(x) = \mathbb{P}(X = x).$$

Definition 17 (Joint Distribution)**.** The **joint distribution** for two discrete random variables X and Y is the collection of values $\{((x, y), \mathbb{P}(X = x, Y = y)) : x \in \mathcal{X}, y \in \mathcal{Y}\}$, where $\mathcal{X}$ is the set of all possible values taken by X and $\mathcal{Y}$ is the set of all possible values taken Y.

Definition 18 (Marginal Distribution)**.** Given the joint distribution for X and Y, the **marginal distribution** for X is as follows:

$$\mathbb{P}(X = x) = \sum_{y \in \mathcal{Y}} \mathbb{P}(X = x, Y = y)$$

Definition 19 (Independence)**.** Random variables X and Y are said to be **independent** if the events $X = x$ and $Y = y$ are independent for all values x, y. Equivalently, the joint distribution of independent R.V's decomposes as

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y), \quad \forall x, y.$$

Definition 20 (Indicator Random variable)**.** $\mathbb{I}_i$, or X_i, denotes the **indicator random variable** that has takes on values $\{0, 1\}$ according to whether a specified event occurs or not. Usually $\{\mathbb{I}_i\}_{i=1}^n$ are mutually independent and they are said to be *independent and identically distributed (i.i.d).*

2.1 Expectation

Definition 21 (Expectation)**.** The **expectation** of a discrete random variable X is defined as

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

Alternatively, we also have

$$\mathbb{E}[X] = \sum_{\omega \in \Omega} X(\omega) \cdot \mathbb{P}(\omega).$$

2.1.1 Linearity of Expectation

Theorem 22 (Linearity of Expectation)**.** For any two random variables X and Y on the same probability space, we have

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

For any constant a, c, we also have

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

Proof. Let $g(X, Y) = X + Y$. Then we have

$$\begin{aligned}
\mathbb{E}[X+Y] &= \sum_{x,y} (x+y)\mathbb{P}(X=x, Y=y) \\
&= \sum_{x,y} x\mathbb{P}(X=x, Y=y) + \sum_{x,y} y\mathbb{P}(X=x, Y=y) \\
&= \sum_{x} \sum_{y} x\mathbb{P}(X=x, Y=y) + \sum_{y} \sum_{x} y\mathbb{P}(X=x, Y=y) \\
&= \sum_{x} x \sum_{y} \mathbb{P}(X=x, Y=y) + \sum_{y} y \sum_{x} \mathbb{P}(X=x, Y=y) \\
&= \sum_{x} x\mathbb{P}(X=x) + \sum_{y} y\mathbb{P}(Y=y) \\
&= \mathbb{E}[X] + \mathbb{E}[Y].
\end{aligned}$$

The proof of the second equality is left as an exercise. □

This is a powerful theorem because this always applies without any assumption about the R.V.s.

Remark. Be careful that this doesn't imply that $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$, or $\mathbb{E}\left[\frac{1}{X}\right] = \frac{1}{\mathbb{E}[X]}$. These are not true in general.

2.2 Variance

Definition 23 (Variance)**.** The **variance** of a random variable X is

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

Definition 24 (Standard Deviation)**.** The **standard deviation** of a random variable X

$$\sigma := \sqrt{\mathrm{Var}(X)}.$$

Theorem 25. For a random variable X,

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

Proof.

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

Note that $\mathbb{E}[X]$ is a constant. □

Fact 26. For any constant c and any random variable X, we have

$$\mathrm{Var}(cX) = c^2\mathrm{Var}(X).$$

Theorem 27. For independent random variables X, Y, we have $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.

Proof.

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

where the second line made crucial use of independence. □

Here's a more general statement:

Theorem 28. If X, Y are independent, then

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

Proof. This is left as an exercise. □

Theorem 29. For **independent** random variables X, Y,

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

Proof.

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

□

2.2.1 Covariance

Covariance is a measure of the joint variability of two random variables.

Definition 30 (Covariance)**.** The **covariance** of random variables X and Y, denoted $\text{Cov}(X, Y)$, is defined as

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

Remark. Some important facts about covariance.

1. If X, Y are independent, then $\text{Cov}(X, Y) = 0$. However, the converse is not true.
2. $\text{Cov}(X, X) = \text{Var}(X)$.
3. *Bilinearity*:
$$\text{Cov}\left(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{m} b_j Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j \text{Cov}(X_i, Y_j).$$
4. For general random variables X and Y,
$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y).$$

2.2.2 Correlation

Definition 31 (Correlation)**.** Suppose X, Y are random variables with $\sigma_X, \sigma_Y > 0$. Then the **correlation** of X and Y is

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

and $-1 \leq \rho(X, Y) \leq 1$.

2.3 Discrete Probability Distribution

2.3.1 Bernoulli Distribution

A **Bernoulli** random variable X, denoted as Bernoulli(p), has a PDF of the form

$$\mathbb{P}(X = i) = \begin{cases} p, & \text{if } i = 1 \\ 1 - p, & \text{if } i = 0, \end{cases}$$

where $0 \leq p \leq 1$.
Expectation:

$$\mathbb{E}[X] = p.$$

Variance:

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

2.3.2 Binomial Distribution

A **binomial** random variable X, denoted as $\text{Bin}(n, p)$, has a PDF of the form

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad \text{for } k = 0, 1, \ldots, n.$$

Quick check on normalization:

$$\sum_{i=0}^{n} \mathbb{P}(X = i) = 1 \implies \sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} = 1.$$

A probabilistic proof of the Binomial Theorem for $a = p$ and $b = 1 - p$.

Fact 32. A binomial random variable is equivalent to sum of n i.i.d Bernoulli variables with parameter p.

Expectation:

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{n} Y_i\right] = \sum_{i=1}^{n} \mathbb{E}[Y_i] = \sum_{i=1}^{n} p = np, \quad \text{where } Y_i \sim \text{Bernoulli}(p).$$

Variance:

$$\text{Var}(X) = \text{Var}\left(\sum_{i=1}^{n} Y_i\right) = np(1-p), \quad \text{where } Y_i \sim \text{Bernoulli}(p).$$

2.3.3 Hypergeometric Distribution

We are given $N = G + B$ balls, where G balls are good and B balls are bad. Sample n balls *without* replacement and observe k successes. Denoted as $\text{Hypergeometric}(N, B, n)$ and has a PDF of the form

$$\mathbb{P}(X = k) = \frac{\binom{G}{k}\binom{B}{n-k}}{\binom{N}{n}}.$$

2.3.4 Geometric Distribution

A **geometric** random variable X, denoted as $\text{Geo}(p)$, has a PDF of the form

$$\mathbb{P}(X = k) = (1-p)^{k-1} p, \quad \text{for } i = 1, 2, 3, \ldots.$$

It represents the number of trials until first success, where p is the probability of success.
Quick check on normalization:

$$\sum_{i=1}^{\infty} \mathbb{P}(X = i) = \sum_{i=1}^{\infty} (1-p)^{i-1} p = p \sum_{i=1}^{\infty} (1-p)^{i-1} = p \cdot \frac{1}{1-(1-p)} = 1.$$

Expectation:

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} \mathbb{P}(X \geq i) = \sum_{x=1}^{\infty} (1-p)^{i-1} = \frac{1}{1-(1-p)} = \frac{1}{p},$$

where the first equality uses the **tail sum formula**, which is on the next page.

Theorem 33 (Tail Sum Formula)**.** Let X be a random variable that takes values in $\{0, 1, 2, \ldots\}$. Then

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

Proof. We can manipulate the formula for the expectation:

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

This is called the *Tail Sum Formula* because we are summing over the tail probabilities of the distribution. □

Remark. Here's a *smarter* way to derive the expectation. Suppose we toss our first coin. There are two possibilities: we get a head with probability p and call it a day, or we get a tail with probability $1 - p$ and we are right back where we just started. In the latter case, we expect $1 + \mathbb{E}[X]$ trials until our first success because we already used one trial. Hence,

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

This makes use of an important property called the **memoryless property**, which will be covered later.

Variance:

$$\text{Var}(X) = \frac{1 - p}{p^2}.$$

2.3.5 Poisson Distribution

A **Poisson** random variable X, denoted as Poisson(λ), has a PDF of the form

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

It is used to model rare events and is an approximation of the limiting case of binomial distribution.

Quick check on normalization:

$$\sum_{i=0}^{\infty} \mathbb{P}(X = i) = \sum_{i=0}^{\infty} \frac{\lambda^k e^{-\lambda}}{k!} = e^{-\lambda} \sum_{i=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda} \cdot e^{\lambda} = 1.$$

Remark. The second equality uses the **Taylor series expansion**

$$e^x = \sum_{i=1}^{\infty} \frac{x^i}{i!}.$$

Expectation:

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{i=0}^{\infty} i \cdot \mathbb{P}(X = i) \\
&= \sum_{i=1}^{\infty} \frac{\lambda^i e^{-\lambda}}{i!} \\
&= \lambda e^{-\lambda} \sum_{i=1}^{\infty} \frac{\lambda^{i-1}}{(i-1)!} \\
&= \lambda e^{-\lambda} e^{\lambda} \quad (e^{\lambda} = \sum_{j=1}^{\infty} \frac{\lambda^j}{j!} \text{ with } j = i-1) \\
&= \lambda.
\end{aligned}$$

Variance:

Similarly, we can calculate $\mathbb{E}[X(X-1)]$ as follows:

$$\begin{aligned}
\mathbb{E}[X(X-1)] &= \sum_{i=0}^{\infty} i(i-1) \cdot \mathbb{P}(X = i) \\
&= \sum_{i=2}^{\infty} i(i-1) \frac{\lambda^i}{i!} \mathrm{e}^{-\lambda} \quad \text{i=0 and i=1 terms are equal to 0)} \\
&= \lambda^2 \mathrm{e}^{-\lambda} \sum_{i=2}^{\infty} \frac{\lambda^{i-2}}{(i-2)!} \\
&= \lambda^2 \mathrm{e}^{-\lambda} \mathrm{e}^{\lambda} \quad (\text{ since } \mathrm{e}^{\lambda} = \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} \text{ with } j = i-2) \\
&= \lambda^2
\end{aligned}$$

Therefore,

$$\mathrm{Var}(X) = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = \mathbb{E}[X(X-1)] + \mathbb{E}[X] - \mathbb{E}[X]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

Theorem 34. Let $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\mu)$ be independent Poisson random variables. Then $X + Y \sim \text{Poisson}(\lambda + \mu)$.

Proof. For all $k = 0, 1, 2, \ldots$, we have

$$\begin{aligned}
\mathbb{P}(X + Y = k) &= \sum_{j=0}^{k} \mathbb{P}(X = j, Y = k - j) \\
&= \sum_{j=0}^{k} \mathbb{P}(X = j)\mathbb{P}(Y = k - j) \\
&= \sum_{j=0}^{k} \frac{\lambda^j}{j!} e^{-\lambda} \frac{\mu^{k-j}}{(k-j)!} e^{-\mu} \\
&= e^{-(\lambda+\mu)} \frac{1}{k!} \sum_{j=0}^{k} \frac{k!}{j!(k-j)!} \lambda^j \mu^{k-j} \\
&= e^{-(\lambda+\mu)} \frac{(\lambda+\mu)^k}{k!}
\end{aligned}$$

where the second equality follows from independence, and the last equality from the binomial theorem. □

Theorem 35. If $X_1, X_2, ..., X_n$ are independent Poisson random variables with parameters $\lambda_1, \lambda_2, \ldots, \lambda_n$ respectively, then

$$X_1 + X_2 + \ldots + X_n \sim \text{Poisson}(\lambda_1 + \lambda_2 + \ldots + \lambda_n).$$

Proof. This can be shown by induction. □

2.4 Conditioning of Random Variables

Definition 36 (Conditional PMF)**.** The **conditional PMF** of a random variable X, conditioned on an event A with $\mathbb{P}(A) > 0$ is defined by

$$p_{X|A}(x) = \mathbb{P}(X = x|A) = \frac{\mathbb{P}(\{X = x\} \cap A)}{\mathbb{P}(A)}.$$

Since

$$\mathbb{P}(A) = \sum_x \mathbb{P}(\{X = x\} \cap A),$$

combining the two gives

$$\sum_x p_{X|A}(x) = 1.$$

Definition 37 (Conditional PMF II)**.** Given two random variables X, Y, the **conditional PMF** of X, conditioned on Y with $Pp_Y(y) > 0$ is defined by

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

We also have

$$\sum_x p_{X|Y}(x|y) = 1.$$

2.4.1 Conditional Expectation

Definition 38 (Conditional Expectation)**.** The **conditional expectation** of X given an event A with $\mathbb{P}(A) > 0$ is defined by

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

For a function $g(X)$, we have

$$\mathbb{E}[g(X)|A] = \sum_x g(x) p_{X|A}(x).$$

Similarly if we condition on a given value y of a random variable Y, then

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

Theorem 39 (Total Expectation)**.** These follow form the **Law of Total Probability**:

- If $A_1, \ldots, A_n$ are disjoint events that form a partition of the sample space, with $\mathbb{P}(A_i) > 0$ for all i, then
$$\mathbb{E}[X] = \sum_x \mathbb{E}[X|A_i]\mathbb{P}(A_i).$$
- Similarly, we have
$$\mathbb{E}[X] = \sum_y p_Y(y)\mathbb{E}[X|Y = y].$$

Proof.

$$\begin{aligned}
\mathbb{E}[X] &= \sum_x x p_X(x) \\
&= \sum_x x \sum_{i=1}^n \mathbb{P}(A_i)\, p_{x|A_i}(x \mid A_i) \\
&= \sum_{i=1}^n \mathbb{P}(A_i) \sum_x x p_{x|A_i}(x \mid A_i) \\
&= \sum_{i=1}^n \mathbb{P}(A_i)\, \mathbb{E}[X \mid A_i]
\end{aligned}$$

The second equality can be verified similarly. □

Definition 40 (Memoryless Property)**.**] A random variable X is memoryless if

$$\mathbb{P}(X > s + t \mid X > s) = \mathbb{P}(X > t).$$

Proof.

$$\begin{aligned}
\mathbb{P}(X = s + t \mid X > s) &= \frac{\mathbb{P}(X > s + t \cap X > s)}{\mathbb{P}(X > s)} \\
&= \frac{\mathbb{P}(X > s + t)}{\mathbb{P}(X > s)} \\
&= \frac{(1-p)^{s+t}}{(1-p)^s} \\
&= (1-p)^t \\
&= \mathbb{P}(X > t).
\end{aligned}$$

□

Fact 41. Geometric random variables are **memoryless**.

Lemma 42.

$$\mathbb{E}[g(X)|X > 1] = \mathbb{E}[g(1+X)].$$

Proof.

$$\begin{aligned}
\mathbb{E}[g(X) \mid X > 1] &= \sum_{k=1}^{\infty} g(k)\mathbb{P}(X = k \mid X > 1) \\
&= \sum_{k=1}^{\infty} g(k)\mathbb{P}(X = k-1) \quad \text{(memoryless property)} \\
&= \sum_{n=1}^{\infty} g(1+n)\mathbb{P}(X = l) \quad (\text{let } n = k-1) \\
&= \mathbb{E}[g(1+X)] \\
&= 1.
\end{aligned}$$

□

Remark. Here's a *clever* trick using memorylessness property to calculate the mean and variance of a geometric random variable. If the first try is successful, we have $X = 1$, and $\mathbb{E}[X|X = 1] = 1$. If the first try fails ($X > 1$), we have wasted one try and we are back where we started. So

$$\mathbb{E}[X|X > 1] = 1 + \mathbb{E}[X].$$

Thus, by total expectation

$$\begin{aligned}
\mathbb{E}[X] &= \mathbb{E}[X|X = 1]\mathbb{P}(X = 1) + \mathbb{P}(X > 1)\mathbb{E}[X|X > 1] \\
&= p + (1-p)(1 + \mathbb{E}[X]),
\end{aligned}$$

from which we obtain

$$\mathbb{E}[X] = \frac{1}{p}.$$

Similar for variance, we have $\mathbb{E}[X^2|X = 1] = 1$ and $\mathbb{E}[X^2|X > 1] = \mathbb{E}[(1+X)^2] = 1 + 2\mathbb{E}[X] + \mathbb{E}[X^2]$ using the lemma proved above. Thus,

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

Using the fact that $\mathbb{E}[X] = \frac{1}{p}$, we have $\mathbb{E}[X^2] = \frac{2}{p^2} - \frac{1}{p}$. Therefore,

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

3 Continuous Probability

3.1 Continuous Random Variables

Definition 43 (Probability Density Function)**.** A **probability density function**, or **PDF**, for a real-valued random variable X is a function $f : \mathbb{R} \to \mathbb{R}$ satisfying:

1. **Non-negativity:** $f(x) \geq 0$ for all $x \in \mathbb{R}$.
2. **Normalization:**
$$\int_{-\infty}^{\infty} f(x)dx = 1.$$

The distribution of X is given by

$$\mathbb{P}(X \in B) = \int_B f_X(x)dx.$$

In particular, the probability that the value of X falls within an interval is

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f(x)dx \quad \text{for all } a < b.$$

For an interval $[x, x + \delta]$ with very small length δ, we have

$$\mathbb{P}([x, x + \delta]) = \int_x^{x+\delta} f_X(t)dt \approx f_X(x) \cdot \delta.$$

3.2 Expectation and Variance

Definition 44 (Expectation)**.** The **expectation** of a continuous random variable X with PDF f is

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

We also have

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

Definition 45 (Variance)**.** The **variance** of a continuous random variable X with PDF f is

$$\text{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \int_{-\infty}^{\infty} x^2 f_X(x)dx - \left(\int_{-\infty}^{\infty} x f_X(x)dx\right)^2.$$

3.3 Continuous Probability Distribution

3.3.1 Exponential Random Variable

An **exponential** random variable X, denoted as $\text{Exp}(\lambda)$, has a PDF of the form

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

Quick check on normalization:

$$\int_{-\infty}^{\infty} f(x)dx = \int_0^{\infty} \lambda e^{-\lambda x}dx = -e^{-\lambda x}\Big|_0^{\infty} = 1.$$

Expectation:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} xf(x)dx = \int_0^{\infty} \lambda x e^{-\lambda x} dx = -xe^{-\lambda x}\Big|_0^{\infty} + \int_0^{\infty} e^{-\lambda x} dx = 0 + \left(-\frac{e^{-\lambda x}}{\lambda}\right)\Big|_0^{\infty} = \frac{1}{\lambda}.$$

Variance:

$$\mathbb{E}\left[X^2\right] = \int_{-\infty}^{\infty} x^2 f(x)dx = \int_0^{\infty} \lambda x^2 e^{-\lambda x} dx = -x^2 e^{-\lambda x}\Big|_0^{\infty} + \int_0^{\infty} 2xe^{-\lambda x} dx = 0 + \frac{2}{\lambda}\mathbb{E}[X] = \frac{2}{\lambda^2}.$$

$$\mathrm{Var}(X) = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

Theorem 46 (Minimum of Exponential Random Variables)**.** Let $X_1, \ldots, X_n$ be independent exponential random variables with parameters $\lambda_1, \ldots, \lambda_n$ respectively. Then the minimum of the random variables is also exponentially distributed:

$$\min\{X_1, \ldots, X_n\} \sim \mathrm{Exp}\left(\lambda_1 + \cdots + \lambda_n\right).$$

Proof.

$$\begin{aligned}
\mathbb{P}\left(\min\{X_1, \ldots, X_n\} > t\right) &= \mathbb{P}\left(X_1 > t, \ldots, X_n > t\right) \\
&= \prod_{i=1}^{n} \mathbb{P}\left(X_i > t\right) \\
&= \prod_{i=1}^{n} e^{-\lambda_i t} \\
&= e^{-\left(\sum_{i=1}^{n} \lambda_i\right)t}.
\end{aligned}$$

□

3.3.2 Cumulative Distribution Functions

Definition 47. For a continuous random variable X, the **cumulative distribution function**, or **CDF**, is the function as follows:

$$F_X(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f(t)dt.$$

It is closely related to the PDF for X:

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

Some properties of a CDF:

- F_X is monotonically nondecreasing:

$$x \leq y \implies F_X(x) \leq F_X(y).$$

-

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

- If X is discrete, the PMF and CDF can be obtained from each other by summing or differencing:

$$F_X(k) = \sum_{i=-\infty}^{k} p_X(i),$$

$$p_X(k) = \mathbb{P}(X \leq k) - \mathbb{P}(X \leq k-1) = F_X(k) - F_X(k-1).$$

3.4 Normal Random Variables

Definition 48 (Normal/Gaussian RV)**.** A **normal** or **Gaussian** random variable X, denoted by $\mathcal{N}(\mu, \sigma^2)$ where μ is the mean and σ^2 is the variance, has a PDF of the form

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

Let's verify that

$$\frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{-(x-\mu)^2/2\sigma^2} dx = 1.$$

Proof. We can show this for $\mu = 0$ and $\sigma^2 = 1$ and this will show for the general case. The trick is to show that

$$\left(\int_{-\infty}^{\infty} f_X(x) dx \right)^2 = 1$$

We have

$$\begin{aligned}
\left(\int_{-\infty}^{\infty} f_X(x) dx \right)^2 &= \left(\int_{-\infty}^{\infty} f_X(x) dx \right) \left(\int_{-\infty}^{\infty} f_Y(y) dy \right) \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_X(x) f_Y(y) dx dy \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{2\pi} e^{-\left(x^2+y^2\right)/2} dx dy.
\end{aligned}$$

Using polar integration, we have $dydx = rdrd\theta$. Then

$$\begin{aligned}
&\int_0^{2\pi} \int_0^{\infty} \frac{1}{2\pi} e^{-r^2/2} r dr d\theta \\
&\quad = \int_0^{\infty} e^{-r^2/2} r dr \\
&\quad = \int_{-\infty}^{0} e^s ds \\
&\quad = 1.
\end{aligned}$$

□

Definition 49 (Standard Normal RV)**.** The PDF of the *standard normal* distribution $\mathcal{N}(0, 1)$ (with mean 0 and variance 1) is

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

Since its CDF cannot be expressed in elementary functions, the CDF is denoted by Φ

$$\Phi(x) = \mathbb{P}(X \leq x) = \mathbb{P}(X < x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt.$$

Remark. The CDF of a normal random variable is symmetrical, so

$$\Phi(-x) = 1 - \Phi(x).$$

Theorem 50 (Normality is preserved by Linear Transformations)**.** If $X \sim \mathcal{N}(\mu, \sigma^2)$ and $a \neq 0, b$ are constants, then the random variable

$$Y = aX + b$$

is also normal. In particular, $Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.

To calculate the CDF for a normal random variable $X \sim \mathcal{N}(\mu, \sigma^2)$, we **standardize** X

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

3.5 Joint PDFs of Multiple Random Variables

Two continuous random variables are **jointly continuous** and can be described in terms of a **joint PDF** f_X, Y that satisfies

$$\mathbb{P}((X,Y) \in B) = \iint_{(x,y)\in B} f_{X,Y}(x,y)dxdy,$$

for every subset B of the 2-dimensional plane. If B is a rectangle of the form $B = \{(x,y)|a \leq b, c \leq y \leq d\}$, we have

$$\mathbb{P}(a \leq X \leq b, c \leq Y \leq d) = \int_c^d \int_a^b f_{X,Y}(x,y)dxdy.$$

3.6 Joint CDFs

The joint CDF of two random variables X and Y is given by

$$F_{X,Y}(x,y) = \mathbb{P}(X \leq x, Y \leq y) = \int_{-\infty}^x \int_{-\infty}^y f_{X,Y}(s,t)dtds.$$

Conversely, the PDF can be recovered from the CDF by differentiating:

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

3.7 Conditioning

Definition 51 (Conditional PDF)**.** For two continuous random variables X, Y with joint PDF $f_{X,Y}$. The **conditional PDF** of X given that $Y = y$ is defined by

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

This is analogous to the discrete case. Similarly for marginalization we have

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

which implies the normalization property

$$\int_{-\infty}^{\infty} f_{X|Y}(x|y)dx = 1.$$

3.7.1 Continuous Bayes' Rule

Similar to the discrete Bayes' Rule, we have

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

combining with the law of total probability gives

$$f_{X|Y}(x|y) = \frac{f_X(x)f_{Y|X}(y|x)}{\int_{-\infty}^{\infty} f_X(t)f_{Y|X}(y|t)dt}.$$

Summary 3.1.
Continuous Uniform Over $[a, b]$:

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

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

Exponential with Parameter λ :

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \ge 0 \\ 0, & \text{otherwise} \end{cases} \qquad F_X(x) = \begin{cases} 1 - e^{-\lambda x}, & \text{if } x \ge 0 \\ 0, & \text{otherwise} \end{cases}$$

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

Normal with Parameters μ **and** $\sigma^2 > 0$:

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

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

4 More on Random Variables

4.1 Derived Distributions

Now let's discuss techniques whereby, given the PDF of X, we calculate the PDF of Y (**derived distribution**).

Theorem 52 (PDF of Linear Function of a Random Variable)**.** Let X and Y be random variables, such that $Y = aX + b$ where $a \neq 0$. Then given f_X, we can derive $f_Y(y)$ as

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right).$$

Proof. We only show for the case where $a > 0$ because the case $a < 0$ is similar. We have

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \leq y) \\ &= \mathbb{P}(aX + b \leq y) \\ &= \mathbb{P}\left(X \leq \frac{y-b}{a}\right) \\ &= F_X\left(\frac{y-b}{a}\right). \end{aligned}$$

Differentiating this with chain rule gives

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \frac{1}{a} f_X\left(\frac{y-b}{a}\right).$$

□

Example 4.1. Suppose $X = \sigma Y + \mu$, where $Y \sim \mathcal{N}(0,1)$. Given that

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

we have

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

which is exactly the PDF for a normal distribution.

Summary 4.1 (Calculation of Derived Distributions)**.** Suppose $Y = g(X)$. Then to find the PDF of Y:

1. Calculate
$$F_Y(y) = \int_{\{x|g(x)\leq y\}} f_X(x)dx.$$
2. Then differentiate
$$f_Y(y) = \frac{dF_Y(y)}{dy}.$$

Example 4.2. Let $Y = X^2$. Then

1. For $y \geq 0$,

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(x^2 \leq y) = \mathbb{P}(-\sqrt{y} \leq x \leq \sqrt{y}) = F_X(\sqrt{y}) - F_X(-\sqrt{y})$$

2. Differentiating gives

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \frac{1}{2\sqrt{y}} f_X(\sqrt{y}) - \frac{1}{2\sqrt{y}} f_X(-\sqrt{y}).$$

4.2 Convolution

Let $Z = X + Y$, where X and Y are both continuous and independent. We are interested in calculating the PDF of Z using law of total probability:

$$f_Z(z) = \int_x f_{X,Z}(x, z).$$

First note that

$$\begin{aligned} F_{Z|X}(z|x) &= \mathbb{P}(X + Y \leq z | X = x) \\ &= \mathbb{P}(Y \leq z - x) \\ &\Rightarrow f_{Z|X}(z|x) = f_Y(z - x). \end{aligned}$$

Now we have

$$f_Z(z) = \int_x f_X(x) f_{Z|X}(z|x) dx = \int_x f_X(x) f_Y(z - x) dx = (f_X * f_Y)(z).$$

$f_Z(z)$ is called the **convolution** of the PDFs of X and Y. We are basically integrating over all possible combinations of X and Y that could sum to z. The discrete case is analogous:

$$\mathbb{P}(Z = z) = \sum_x \mathbb{P}(X = x)\mathbb{P}(Y = z - x).$$

4.3 Law of Iterated Expectations

Theorem 53 (Law of Iterated Expectations)**.**

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

Proof.

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

□

Property 1.

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

This follows from the fact that $g(Y)$ is a constant.

Example 4.3. We have a biased coin with probability of head is Y, which is random over $[0, 1]$. Let X be the number of heads obtained. What is $\mathbb{E}[X]$?

Solution.

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[nY] = n\mathbb{E}[Y] = \frac{n}{2}.$$

4.4 Law of Total Variance

Theorem 54 (Law of Total Variance)**.**

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

Proof. Using the law of iterated expectation, we have

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

□

Example 4.4. Using the previous example, we now compute the variance of X.

Solution.

$$\begin{aligned}
\mathrm{Var}(X) &= \mathbb{E}[\mathrm{Var}(X|Y)] + \mathrm{Var}(\mathbb{E}[X|Y]) \\
&= \mathbb{E}[nY(1-Y)] + \mathrm{Var}(nY) \\
&= n(\mathbb{E}[Y] - \mathbb{E}[Y^2]) + n^2\mathrm{Var}(Y) \\
&= \frac{n}{6} + \frac{n^2}{12}.
\end{aligned}$$

4.5 Order Statistics

Definition 55 (Order Statistics)**.** Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. random variables with common density $f_X(x)$ and CDF $F_X(x)$. We define $X^{(i)} = x^{(i)}$ the ith **order statistic**, which is the ith smallest element of the set. Then $X^{(1)}$ is the minimum while $X^{(n)}$ is the maximum.

Theorem 56. If X has pdf f_X, the marginal PDF of the i^{th} order statistic is

$$f_{X^{(i)}}(x) = n\binom{n-1}{i-1}\left(F_X(x)\right)^{i-1}\left(1-F_X(x)\right)^{n-i} f_X(x).$$

Proof. By definition,

$$\mathbb{P}(X^{(i)} \in (x, x+dx)) \approx f_{X^{(i)}}(x)dx.$$

In order for the ith smallest point to lie between x and $x+dx$, we need

1. $i-1$ points must lie in the interval $(-\infty, x)$.
2. One point must lie in $(x, x+dx)$.
3. $n-i$ values to lie in the interval $(x+dx, \infty)$.

We have n choices for a point, and $\binom{n-1}{i-1}$ choices to distribute the rest, so we have $n\binom{n-1}{i-1}$ ways to distribute the points. Then combining these gives

$$f_{X^{(i)}}(x)dx = n\binom{n-1}{i-1}\left(F_X(x)\right)^{i-1}\left(1-F_X(x)\right)^{n-i} f_X(x)dx.$$

□

Example 4.5 (Beta Distribution)**.** Suppose $X \sim U[0,1]$, where $f_X(x) = 1$ and $F_X(x) = x$ for $0 \le x \le 1$. Then the ith order statistic for X is

$$f_X^{(i)}(x) = n\binom{n-1}{i-1}x^{i-1}(1-x)^{n-i}.$$

This is a special case of a **Beta Distribution**.

Example 4.6. What is the probability that the 9th smallest out of 10 drawings from $X \sim U[0,1]$ is greater than 0.8?

Solution.

$$f_{X^{(9)}}(x) = \frac{10!}{8!1!}x^8(1-x) = 10x^8 - 90x^9 \implies \mathbb{P}(X^{(9)} > 0.8) = \int_{0.8}^{1}(90x^8 - 90x^9)dx.$$

5 Moment Generating Functions (Transforms)

Definition 57 (Moment Generating Function)**.** The **transform** (or MGF) of a random variable X is a function $M_X(s)$ of a scalar parameter s, defined by

$$M_X(s) = \mathbb{E}\left[e^{sX}\right].$$

Discrete case:

$$M_X(s) = \sum_x e^{sx} p_X(x).$$

Continuous case:

$$M_X(s) = \int_{-\infty}^{\infty} e^{sx} f_X(x) dx.$$

Recall the Taylor series expansion for e:

$$e^{sX} = 1 + sX + \frac{s^2}{2!}X^2 + \frac{s^3}{3!}X^3 + \cdots.$$

Then we have

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

Theorem 58. We can use MGF to compute the *moments* of X as follows:

$$\frac{d^n}{ds^n}[M_X(s)]\bigg|_{s=0} = \int_{-\infty}^{\infty} x^n f_X(s) dx = \mathbb{E}[X^n].$$

Remark. Convolution becomes multiplication in the MGF domain and so MGFs simplify computations for us and it can be used to prove the Central Limit Theorem as we will soon cover.

Property 2 (Properties of MGFs)**.** $M_X(s)$ satisfies the following properties:

1. $M_X(0) = 1$.
2. For $Y = aX + b$, $M_Y(s) = e^{sb} M_X(as)$.
3. If $X > 0$, then $M_X(-\infty) = 0$.
4. If $X < 0$, then $M_X(\infty) = 0$.

Proof. For the first one, we have

$$M_X(0) = \mathbb{E}[e^{0X}] = 1.$$

For the second one, we have

$$M_Y(s) = \mathbb{E}[e^{sY}] = \mathbb{E}[e^{s(aX+b)}] = e^{sb}\mathbb{E}[e^{asX}] = e^{sb} M_X(as).$$

The rest will be exercise. □

Example 5.1 (Exponential MGF). Suppose $X \sim \text{Exp}(\lambda)$. Then

$$M_X(s) = \mathbb{E}[e^{sX}] = \int_0^\infty e^{sx} f_X(x)dx = \lambda \int_0^\infty e^{-\lambda x} e^{sx} dx = \lambda \frac{e^{-(\lambda - s)x}}{-(\lambda - s)}\Big|_0^\infty = \frac{\lambda}{\lambda - s}$$

where $s < \lambda$ must hold for the integral to converge. Then we obtain $\mathbb{E}[X] = M_X'(0) = \frac{\lambda}{(\lambda - s)^2}\Big|_{s=0} = \frac{1}{\lambda}$, and $\mathbb{E}[X^2] = M_X''(0) = \frac{2}{\lambda}$.

Example 5.2 (Poisson MGF). Suppose $X \sim \text{Poisson}(\lambda)$. Then

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

From this we get $M_X'(0) = \lambda$ and $M_X''(0) = \lambda^2 + \lambda$.

Example 5.3 (Normal MGF). Suppose $X \sim \mathcal{N}(0, 1)$. Then

$$\begin{aligned}
\mathbb{E}\left[e^{sX}\right] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty e^{sx} e^{-x^2/2} dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty e^{sx - x^2/2} dx \\
&= \frac{e^{s^2/2}}{\sqrt{2\pi}} \int_{-\infty}^\infty e^{-\left(x^2/2 - sx + s^2/2\right)} \\
&= e^{s^2/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty e^{-(x-s)^2/2} \\
&= e^{s^2/2}.
\end{aligned}$$

The third line is simply completing the square in the exponent, and the last line uses the fact that $\frac{1}{\sqrt{2\pi}} e^{-(x-s)^2/2}$ is the PDF of a standard normal that has been shifted by s, and so must integrate to 1. It is left as an exercise to verifty that $M_X'(0) = 0$ and $M_X''(0) = 1$.

Remark. If $Y \sim \mathcal{N}\left(\mu, \sigma^2\right)$, then $Y = \sigma X + \mu$ and we have

$$\mathbb{E}\left[e^{sY}\right] = \mathbb{E}\left[e^{s(\sigma X + \mu)}\right] = e^{\mu s} \mathbb{E}\left[e^{\sigma Y s}\right] = e^{\mu s + \sigma^2 s^2/2}.$$

5.1 Inversions of transforms

The transform $M_X(s)$ is invertible, i.e. it can be used to determine the probability law of the random variable X.

Property 3 (Inversion Property). The transform $M_X(s)$ associated with a random variable X uniquely determines the CDF of X, assuming that $M_X(s)$ is finite for all s in some interval $[-a, a]$, where a is a positive number.

Remark. In practice, transforms are usually inverted by "pattern matching," based on tables of known distribution-transform pairs.

Example 5.4. Suppose we have an MGF of $M_X(s) = \frac{1}{2}e^{-3s} + \frac{1}{4}e^{200s} + \frac{1}{4}e^{s}$. Recall that

$$M_X(s) = \sum_x e^{sx} p_X(x),$$

we can recover our PDF as

$$P(X = k) = \begin{cases} 1/2 & \text{when } k = -3 \\ 1/4 & \text{when } k = 200 \\ 1/4 & \text{when } k = 1. \end{cases}$$

5.2 Sums of Independent Random Variables

Transform methods are especially convenient when dealing with a sum of random variables. This is because addition of independent random variables corresponds to multiplication of transforms (providing a nice alternative to the convolution formula).

Suppose X and Y are independent random variables, and let $Z = X + Y$. The transform associated with Z is

$$M_Z(s) = \mathbb{E}\left[e^{sZ}\right] = \mathbb{E}\left[e^{s(X+Y)}\right] = \mathbb{E}\left[e^{sX}e^{sY}\right] = M_X(s)M_Y(s).$$

By the same argument, if $X_1, \ldots, X_n$ is a collection of independent random variables and $Z = \sum_{i=1}^n X_i$, then

$$M_Z(s) = \prod_{i=1}^n M_{X_i}(s), \implies \mathbb{E}\left[\exp\left(s\sum_{i=1}^n X_i\right)\right] = \prod_{i=1}^n \mathbb{E}[e^{sX_i}].$$

Remark. In summary, the MGF of the sum of two random variables is the product of their MGFs.

Example 5.5 (Sum of Independent Normal Random Variables is Normal)**.** Suppose $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \mu_Y^2)$. Then

$$M_Z(s) = M_X(s)M_Y(s) = \exp\left(\left(\frac{\sigma_X^2 + \sigma_Y^2}{2}\right)s^2 + (\mu_X + \mu_Y)s\right),$$

which is the transform associated with a normal random variable $\mathcal{N}(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$.

Summary 5.1 (Transforms for Common Random Variables).
$X \sim$**Bernoulli**(p)
PMF:

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

MGF:

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

$X \sim$ **Binomial**(n, p)
PMF:

$$p_X(x) = \binom{n}{x} p^k (1-p)^{n-x},$$

MGF:

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

$X \sim$ **Geometric**(p)
PMF:

$$p_X(k) = p(1-p)^{x-1},$$

MGF:

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

$X \sim$ **Poisson**(λ)
PMF:

$$p_X(x) = \frac{e^{-\lambda}\lambda^x}{x!},$$

MGF:

$$M_X(s) = e^{\lambda(e^s-1)}.$$

$X \sim$ **Uniform**$[a, b]$
PMF:

$$p_X(k) = \frac{1}{b-a+1},$$

MGF:

$$M_X(s) = \frac{e^{sa}\left(e^{s(b-a+1)}-1\right)}{(b-a+1)\left(e^s-1\right)}.$$

PDF:

$$f_X(x) = \frac{1}{b-a},$$

MGF:

$$M_X(s) = \frac{e^{sb}-e^{sa}}{s(b-a)}.$$

$X \sim$ **Exponential**(λ)
PDF:

$$f_X(x) = \lambda e^{-\lambda x},$$

MGF:

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

$X \sim$ **Normal** $\left(\mu, \sigma^2\right)$
PDF:

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

MGF:

$$M_X(s) = e^{\left(\sigma^2 s^2/2\right)+\mu s}.$$

6 Limit Theorems

6.1 Markov's Inequality

Theorem 59 (Markov's Inequality)**.** For a **non-negative** random variable X with finite mean,

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

for any positive constant c.

Proof.

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}[X|X \leq c]\mathbb{P}(X \leq c) + \mathbb{E}[X|X \geq c]\mathbb{P}(X \geq c) \\ &\geq 0 + c\mathbb{P}(X \geq c).\end{aligned}$$

We lower bound by 0 since X is nonnegative and we lowerbound $\mathbb{E}[X|X \geq c]$ by c since we're conditioning on $X \geq t$. □

6.2 Chebyshev's Inequality

We have seen that the variance (or, more correctly the standard deviation) is a measure of *spread*, or deviation from the mean. We can now make this intuition quantitatively precise:

Theorem 60 (Chebyshev's Inequality)**.** For a random variable X with finite expectation $\mathbb{E}[X] = \mu$,

$$\mathbb{P}[|X - \mu| \geq c] \leq \frac{\text{Var}(X)}{c^2}$$

and for any positive constant c.

Proof. Define $Y = (X-\mu)^2$ and note that $\mathbb{E}[Y] = \mathbb{E}\left[(X-\mu)^2\right] = \text{Var}(X)$. Also, notice that the event that we are interested in, $|X-\mu| \geq c$, is exactly the same as the event $Y = (X-\mu)^2 \geq c^2$. Therefore, $\mathbb{P}[|X-\mu| \geq c] = \mathbb{P}\left[Y \geq c^2\right]$. Moreover, Y is obviously nonnegative, so we can apply Markov's inequality to get

$$\mathbb{P}[|X-\mu| \geq c] = \mathbb{P}\left[Y \geq c^2\right] \leq \frac{\mathbb{E}[Y]}{c^2} = \frac{\text{Var}(X)}{c^2}.$$

□

Remark. Chebyshev's bound is tighter than Markov's bound.

6.3 Chernoff Bounds

Theorem 61 (Chernoff Bound)**.** For all $s > 0$,

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[e^{sX}]}{e^{sa}} \leq \min_s \left(e^{-sa}\mathbb{E}\left[e^{sX}\right]\right)$$

Proof. Since $s > 0$ and e^x is monotonic, using Markov's inequality we have

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

□

Remark. We can optimize over s to get the tightest bound (set derivative to zero). Although Chernoff's uses all moments, it is not guaranteed that it is better than Markov's or Chebyshev's.

Example 6.1. Suppose $X \sim \mathcal{N}(0, \sigma^2)$. We have

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[e^{-sX}]}{e^{sa}} = \frac{e^{-s^2\sigma^2/2}}{e^{sa}}.$$

Optimizing over s gives

$$\arg\min_s \frac{e^{-s^2\sigma^2/2}}{e^{sa}} = \arg\min_s \frac{s^2\sigma^2}{2} - sa \implies s = t/\sigma^2,$$

which gives $\mathbb{P}(X \geq a) \leq e^{-a^2/2\sigma^2}$.

6.4 Weak Law of Large Numbers

The **Weak Law of Large Numbers** asserts that the sample mean of a large number of i.i.d. random variables is very close to the true mean, with high probability. If $X_1, \ldots, X_n$ is a sequence of i.i.d. random variables with mean μ and variance σ^2, then the sample mean is

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

The expectation is

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

and the variance of is

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

Applying the Chebyshev's inequality, we have

$$\mathbb{P}(|M_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2} \quad \text{for any } \epsilon > 0.$$

As n increases, the Chebyshev's bound goes to 0! As a consequence, we obtain the formal WLLN:

Theorem 62 (Weak Law of Large Numbers)**.** Let $X_1, \ldots, X_n$ be i.i.d. random variables with mean μ. For every $\epsilon > 0$, we have

$$\mathbb{P}(|M_n - \mu| \geq \epsilon) = \mathbb{P}\left(\left|\frac{X_1 + \ldots + X_n}{n} - \mu\right| \geq \epsilon\right) \to 0 \quad \text{as } n \to \infty.$$

Remark. The WLLN suggests that for large n, the bulk of the distribution of M_n is concentrated near μ. Essentially, the sample mean should converge to the true mean.

6.5 Convergence in Probability

Definition 63 (Convergence of a Deterministic Sequence)**.** Let $a_1, \ldots, a_n$ be a sequence of real numbers, and let a be another real number. We say that the sequence a_n **converges** to a, or $\lim_{n\to\infty} a_n = a$, if for every $\epsilon > 0$ there exists some n_0 such that

$$|a_n - a| \leq \epsilon \text{ for all } n \geq n_0.$$

Definition 64 (Convergence in Probability)**.** Let $X_1, \ldots, X_n$ be a sequence of random variables (not necessarily independent), and let a be a real number. We say that the sequence X_n **converges to a in probability** if, for every $\epsilon > 0$, we have

$$\lim_{n\to\infty} \mathbb{P}\left(|Y_n - a| \geq \epsilon\right) = 0.$$

In other words, for every $\epsilon > 0$ and every $\delta > 0$, there exists some n_0 such that

$$P\left(|Y_n - a| \geq \epsilon\right) \leq \delta \quad \text{for all } n \geq n_0$$

If we call ϵ the *accuracy* level and δ the *confidence* level, then X_n can be equal to a within any level of accuracy and confidence provided n is sufficiently large.

Example 6.2. $X_1, \ldots, X_n$ are i.i.d. $U[-1, 1]$. If $Y_n = X_n/n$, does Y_n converge in probability? To what?

Solution. Y_n should converge in probability to 0, since X_n is something between -1 and 1 while n will only get larger and larger. We have

$$F_n(y) = F_X(ny) \Longrightarrow f_{Y_n}(y) = nf_X(ny)$$

Then $\mathbb{P}\left(|Y_n - 0| > \epsilon\right) = 0$ if $\frac{1}{n} < \epsilon$, or $n > \frac{1}{\epsilon}$.

Example 6.3. If $X_1, \ldots, X_n$ are i.i.d. $U[0, 1]$ and $Y_n = \min\left(X_1, \ldots, X_n\right)$, then

$$\begin{aligned} P\left(|Y_n - 0| > \epsilon\right) &= \prod_{i=1}^{n} P\left(X_i > \epsilon\right) \\ &= (1 - \epsilon)^n \to 0 \text{ as } n \to \infty, \end{aligned}$$

which intuitively makes sense.

Example 6.4. Suppose we have an arrival process where we divide the number line into exponentially increasing sized intervals:

$$I_k = \left\{2^k, 2^k + 1, \ldots, 2^{k+1} - 1\right\}$$

And suppose we have exactly one arrival in each interval. So we let $Y_n = 1$ if there is an arrival at time n, and $Y_n = 0$ if there is no arrivals. We then have that

$$\begin{gathered}\mathbb{P}(Y_1 = 1) = 1 \\ \mathbb{P}(Y_2 = 1) = \mathbb{P}(Y_3 = 1) = 1/2 \\ \mathbb{P}(Y_n = 1) = \tfrac{1}{2^k} \quad \text{if } n \in I_k\end{gathered}$$

This implies that

$$\lim_{n\to\infty} \mathbb{P}(|Y_n - 0| \geq \epsilon) = \lim_{n\to\infty} \mathbb{P}(Y_n = 1) = \lim_{k\to\infty} \frac{1}{2^k} = 0,$$

which means that Y_n converges in probability to 0 .

Remark. The above example highlights the weakness of WLLN. We can see of course that for any finite n, there are certainly an infinite number of 1's after n, yet it still converges in probability.

Definition 65 (Convergence with Probability 1)**.** Let $X_1, X_2, \ldots$ be a sequence of random variables (not necessarily independent). We say that X_n **converges to** c **with probability** 1 (or **almost surely**) if

$$\mathbb{P}\left(\lim_{n\to\infty} X_n = c\right) = 1,$$

also denoted by $X_n \xrightarrow{\text{a.s.}} c$.

Remark. Convergence with probability 1 implies convergence in probability, but the converse is not necessarily true.

Theorem 66 (Strong Law of Large Numbers)**.** Let $(X_n)_{n\geq 1}$ be a sequence of i.i.d. random variables with mean μ. Then, the sequence of sample means $M_n = \sum_{i=1}^n X_i/n$ converges to μ with probability 1, in the sense that

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} = \mu\right) = 1.$$

Remark. The difference between WLLN and SLLN is that WLLN states that the probability $\mathbb{P}(|M_n - \mu| \geq \epsilon)$ of a significant deviation of M_n from μ goes to zero as $n \to \infty$. However, for any finite n, this probability can be positive and it is conceivable that once in a while, even if infrequently, M_n deviates significantly from μ. The WLLN provides no conclusive information on the number of such deviations. On the other hand, the SLLN states that with probability 1 M_n converges to μ. This implies that for any given $\epsilon > 0$, the probability that the difference $|M_n - \mu|$ will exceed ϵ an infinite number of times is equal to zero.

6.6 The Central Limit Theorem

Let $X_1, \ldots, X_n$ be a sequence of i.i.d. random variables with mean μ and variance σ^2. We define

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}}$$

We can see that

$$\mathbb{E}[Z_n] = \frac{\sum_{i=1}^n \mathbb{E}[X_i] - n\mu}{\sigma\sqrt{n}} = 0$$

and

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

which brings us to the central limit theorem:

Theorem 67 (Central Limit Theorem)**.** Let $X_1, \ldots, X_n$ be a sequence of i.i.d. random variables with common mean μ and variance σ^2, and define

$$Z_n = \frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}}.$$

Then the CDF of Z_n converges to the standard normal CDF

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

in the sense that

$$\lim_{n\to\infty} P(Z_n \le z) = \Phi(z) \quad \text{for every } z.$$

Proof. If $Y \sim \mathcal{N}(0,1)$, $M_Y(s) = \mathbb{E}[e^{sY}] = e^{s^2/2}$. Then WLOG suppose $X_1, X_2, \ldots, X_n$ are i.i.d. with mean 0 and variance 1. Let

$$Z = \frac{X_1 + X_2 + \cdots + X_n}{\sqrt{n}} \implies \mathbb{E}[Z] = 0, \mathrm{Var}(Z) = 1.$$

Then

$$\begin{aligned} M_Z(s) = \mathbb{E}[e^{sZ}] &= \mathbb{E}\left[\exp\left(\frac{s}{\sqrt{n}}\sum_{i=1}^n X_i\right)\right] \\ &= \prod_{i=1}^n \mathbb{E}\left[e^{\frac{sX_i}{\sqrt{n}}}\right] \\ &= \mathbb{E}\left[e^{\frac{sX_i}{\sqrt{n}}}\right]^n \\ &= \left[M_X\left(\frac{s}{\sqrt{n}}\right)\right]^n \end{aligned}$$

Recall Taylor's theorem: any infinitely differentiable function can be written as $f(x) = f(a) + f'(a)(x-a) + \cdots + f^{(n)}(a)(x-a)^n + \cdots$. Then we have

$$\begin{aligned} M_X(s) &= M_X(0) + M_X'(0)s + M_X''(0)\frac{s^2}{2!} + M_X'''(0)\frac{s^3}{3!} + \ldots \\ &= 1 + \mathbb{E}[X]s + \mathbb{E}[X^2]\frac{s^2}{2} + \mathbb{E}[X^3]\frac{s^3}{6} + \ldots \\ &= 1 + \frac{1}{2}s^2 + \frac{s^3}{6}\mathbb{E}[X^3] + \ldots \end{aligned}$$

Hence, we get

$$M_Z(s) = \left[M_X\left(\frac{s}{\sqrt{n}}\right)\right]^n = \left[1 + \frac{s^2}{2n} + \frac{s^3}{6n^{3/2}}\mathbb{E}[X^3] + \dots\right]^n$$

$$\implies \lim_{n\to\infty} M_Z(s) = \lim_{n\to\infty}\left[1 + \frac{s^2}{2n} + \frac{s^3}{6n^{3/2}}\mathbb{E}[X^3] + \dots\right]^n,$$

which resembles the form $\lim_{n\to\infty}(1+\frac{x}{n})^n = e^x$. We write can rewrite this as

$$\lim_{n\to\infty} M_Z(s) = \lim_{n\to\infty}\left[1 + \frac{s^2/2}{n} + O\left(\frac{1}{n}\right)\right]^n = e^{s^2/2} = M_Y(s),$$

thus we have shown that Z converges in distribution to $\mathcal{N}(0,1)$. □

Example 6.5 (Polling)**.** Suppose there are n randomly sampled voters who indicate if they support candidate X. So $X_i = 1$ if yes, and $X_i = 0$ otherwise. Suppose we want a 95% confidence interval that $|M_n - p| < \epsilon$ where p is the true probability that each voter supports the candidate, and $M_n = \frac{1}{n}\sum X_i$ is the empirical mean. Chebyshev's states that

$$\mathbb{P}\left(|M_n - p| \geq a\right) \leq \frac{\text{Var}\left(M_n\right)}{a^2}.$$

Since $\text{Var}\left(X_i\right) = p(1-p) \leq 1/4$, $\text{Var}\left(M_n\right) = \frac{1}{n}\text{Var}\left(X_i\right) \leq \frac{1}{4n}$ Suppose we want to know the p value to within 0.1 with probability at least 95%. In math terms, we need

$$\mathbb{P}\left(|M_n - p| \geq 0.1\right) \leq 0.05$$

and we have

$$\mathbb{P}\left(|M_n - p| \geq 0.1\right) \leq \frac{\text{Var}\left(M_n\right)}{0.1^2} \leq \frac{1}{4n(0.01)} \implies n \geq 500.$$

On the other side, CLT states that

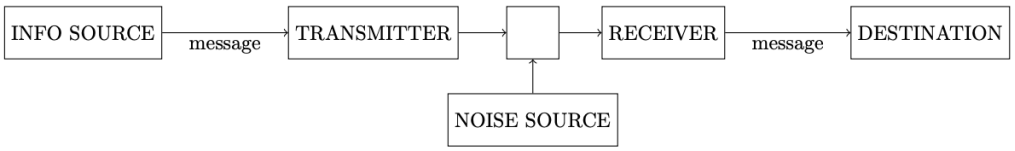
$$\frac{M_n - \mathbb{E}\left[M_n\right]}{\sqrt{\text{Var}\left(M_n\right)}} \to \mathcal{N}(0,1) \implies \mathbb{P}\left(\frac{|M_n - p|}{1/2\sqrt{n}} \geq \frac{0.1}{2\sqrt{n}}\right) \leq 0.05.$$

Since this is roughly a standard normal, we use the fact that 95% of the probability mass lies within 2 or 1.96 standard deviations. So we have

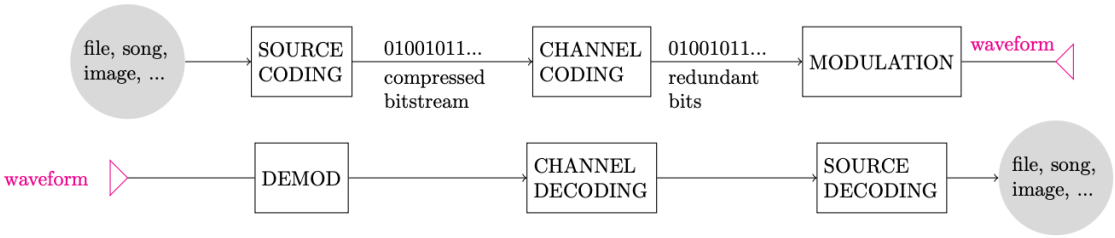
$$0.2\sqrt{n} \geq 2 \Rightarrow n \geq 100.$$

Remark. CLT provides a tighter bound that Chebyshev's.

7 Information Theory



Block diagram of communication system.



A more detailed diagram of a communication channel.