PROBABILITY REVIEW

1. Probability

1.1. Sample Space, Events and Probabilities.

Definition 1.1. The **sample space** Ω is the set of all possible outcomes of a random experiment. Points $\omega \in \Omega$ are called **sample outcomes** or **elements**. Subsets of Ω are called **events.** Ω and ϕ are called **true event** and **null event**, respectively.

Example 1.2. If we toss a coin twice then the sample space $\Omega = \{HH, HT, TH, TT\}$. The event that "at least one head appears" is $A = \{HH, HT, TH\}$.

The sample space in the above example is discrete, and the number of elements $|\Omega|$ is finite. We can also have countable infinite sample space or continuous (uncountable) sample space.

Example 1.3.
(1) If we toss a coin until we see the first head, then the sample space $\Omega = \{H, TH, TTH, TTTH, \ldots\}$ is countable infinite.
(2) Let ω be the waiting time for the next bus. Then $\Omega = (0, \infty)$. The event that "next bus comes in less than 5 minutes" is $A = (0, 5)$. Note that it usually does not hurt to make Ω larger than needed.

Sometimes the sample space can be a mixture of discrete and continuous elements.

Example 1.4. In a random experiment we first toss a coin, and if it is head we randomly choose a number from $\{1, 2, \ldots, 6\}$, otherwise randomly select a real number from $[0, 1]$. Then the sample space (the set of all outcomes) $\Omega = \{1, 2, \ldots, 6\} \cup [0, 1]$.

Since events are subsets of Ω, we need to review some set operations. Given events A, B and A_i $(i = 1, 2, \ldots)$:

- $A^c = \{\omega \in \Omega : \omega \notin A\}$ is the complement of A;
- $A \cup B = \{\omega \in \Omega : \omega \in A \text{ or } \omega \in B\}$ is the event that either A or B occurs;
- $A \cap B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \in B\}$ is the event that both A and B occur (also denoted as AB);
- $A - B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \notin B\}$ is the event that A occurs and B does not occur;
- If for any $\omega \in A$ we have $\omega \in B$ as well, then we denote $A \subset B$. In other words, A is a subset of B.
- $\cup_{i=1}^{\infty} A_i = \{\omega \in \Omega : \omega \in A_i \text{ for at least one } i\}$;
- $\cap_{i=1}^{\infty} A_i = \{\omega \in \Omega : \omega \in A_i \text{ for all } i\}$;
- $A_1, A_2, \ldots$ are **disjoint** or **mutually exclusive** if $A_i \cap A_j = \phi$ for all $i \neq j$.
- A **partition** of Ω is a sequence of disjoint sets $A_1, A_2, \ldots$ such that $\cup_{i=1}^{\infty} A_i = \Omega$.
- A sequence of sets $A_1, A_2, \ldots$ is monotone increasing if $A_1 \subset A_2 \subset A_3 \subset \ldots$ and we define $A = \lim_{n\to\infty} A_n = \cup_{i=1}^{\infty} A_i$; A sequence of sets $A_1, A_2, \ldots$ is monotone decreasing if $A_1 \supset A_2 \supset \ldots$ and we define $A = \lim_{n\to\infty} A_n = \cap_{i=1}^{\infty} A_i$. The former can be written as $A_n \uparrow A$ and the latter can be written as $A_n \downarrow A$, and either case can be written as $A_n \to A$.

Example 1.5. Let $\Omega = \mathbb{R}$ and let $A_i = [0, 1/i)$ for $i = 1, 2, \ldots$. Then $A_1, A_2, \ldots$ are monotone decreasing and $\cup_{i=1}^{\infty} A_i = [0, 1)$ and $\cap_{i=1}^{\infty} A_i = \{0\}$. If instead we define $A_i = (0, 1/i)$ then we have $\cup_{i=1}^{\infty} A_i = (0, 1)$ and $\cap_{i=1}^{\infty} A_i = \phi$.

Definition 1.6. Given a set (an event) A, the **indicator function** of A is defined as

$$I_A(\omega) = I(\omega \in A) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A \end{cases}.$$

Date: August 1, 2006.
PLEASE LET ME KNOW IF YOU FIND ANY ERROR IN THE NOTES.

Note. Later we will see that indicator function can help us understand the connection between probability and expectation. It can also be useful in proving some inequalities.

We want to assign a real number $\mathbb{P}(A)$ to every event A[1] so that it can be used to measure the "volume or size" of the event.

Definition 1.7. A **probability measure** or **probability distribution** is a real-valued function on events $A \subset \Omega$ that satisfies the following three axioms:

(1) $\mathbb{P}(A) \geq 0$ for every A
(2) $\mathbb{P}(\Omega) = 1$
(3) If $A_1, A_2, \ldots$ is a sequence of mutually exclusive events then

$$\mathbb{P}(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

Note. The first axiom specifies that $\mathbb{P}(A)$ is nonnegative; the second axiom defines the probability of the true event Ω to be 1; and the last axiom is about "countable additivity". Also note that countable additivity implies finite additivity: if $A_1, A_2, \ldots, A_n$ are disjoint, then $\mathbb{P}(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} \mathbb{P}(A_i)$.

Theorem 1.8. *(Properties of Probability)*

(1) $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$ (and thus $\mathbb{P}(\phi) = 0$)
(2) $0 \leq \mathbb{P}(A) \leq 1$
(3) If $A \cap B = \phi$ then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$
(4) For any two events A and B, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(AB)$.

Proof. We only prove (4):

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

□

Example 1.9. If we toss a die twice, then the sample space $\Omega = \{(i, j) : i, j \in \{1, 2, \ldots, 6\}\}$. If we further assume that the die is fair and each outcome is equally likely, then $\mathbb{P}(A) = |A|/36$ where $|A|$ denotes the number of elements in A. For example, if A is the event that the sum of the dice is greater than 10, then $\mathbb{P}(A) = 3/36 = 1/12$.

Note. Ω in the above example is called a **uniform probability distribution**, due to the fact that each outcome is equally likely.

1.2. Independence and Conditional Probability.

Definition 1.10. Two events A and B are **independent** if $\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B)$. A set of events $\{A_i : i \in I\}$ is independent if $\mathbb{P}(\cap_{j \in J} A_j) = \prod_{j \in J} \mathbb{P}(A_j)$ for every finite subset $J \subset I$.

Note. There is also "pairwise independent" which is weaker. A set of events $\{A_i : i \in I\}$ is said to be pairwise independent if every pair of events $A_i, A_j (i \neq j)$ is independent.

Intuitively, if A and B are independent, then whether A happens or not does not affect the likelihood of B occurring. Suppose two events A and B with positive probability ($\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$) that are disjoint, then they cannot be independent (prove it). Independence can be used to simplify computation, as shown in the following example.

[1] Technically speaking, not every event can be assigned a probability. We only assign probabilities to sets in a σ-field.

Example 1.11. Flip a fair coin 10 times. Let A be the event that at least one head occurs, and let B_i be the event that the j-th toss results in a tail. Then

$$\begin{aligned}\mathbb{P}(A) &= 1-\mathbb{P}(A^c)\\ &= 1-\mathbb{P}(B_1B_2\dots B_{10})\\ &\overset{\text{by independence}}{=} 1-\mathbb{P}(B_1)\mathbb{P}(B_2)\dots\mathbb{P}(B_{10})\\ &= 1-2^{-10}.\end{aligned}$$

Definition 1.12. If $\mathbb{P}(B)>0$ then the **conditional probability** of A given B is

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

Think of $\mathbb{P}(A|B)$ as the fraction of times A occurs among those in which B occurs. Note that (1) for any two events A and B we have $\mathbb{P}(AB)=\mathbb{P}(A|B)\mathbb{P}(B)=\mathbb{P}(B|A)\mathbb{P}(A)$; (2) if events A and B are independent, then we have $\mathbb{P}(A|B)=\mathbb{P}(A)$.

Theorem 1.13. *(The Law of Total Probability) Let $A_1,A_2,\dots,A_n$ be a partition of Ω. Then for any event B we have*

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

Proof. Since $BA_1,BA_2,\dots,BA_n$ is a partition of B, we have

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

□

Theorem 1.14. *(Bayes Theorem) Let $A_1,\dots,A_n$ be a partition of Ω such that $\mathbb{P}(A_i)>0$ for each i. If $\mathbb{P}(B)>0$ then, for each $i=1,\dots,n$,*

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

Proof. By the definition of conditional probability we have

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

□

2. Random Variables

2.1. Distribution and Probability Functions.

Definition 2.1. A random variable is a mapping[3] $X:\Omega\to\mathbb{R}$ that assigns a real number $X(\omega)$ to each outcome ω.

After random variables are introduced, we often work directly with them and not mention the sample space any more. However, it is important to keep in mind that any random variable is associated with some underlying sample space.

Example 2.2. Toss a die twice, and let $X(\omega)$ be the sum of the dice. For example, if $\omega=(1,5)$ then $X(\omega)=6$. For continuous sample space, let $\Omega=\{(x,y):x^2+y^2\le 1\}$ be the unit disk. Any outcome ω can be written in the form of $\omega=(x,y)$. Some examples of random variables are $X(\omega)=x$, $Y(\omega)=y$, $Z(\omega)=x^2y$, etc.

[3]Technically speaking, a random variable must be a measurable function.

Given a random variable X and a subset A of the real line, define $X^{-1}(A) = \{\omega \in \Omega : X(\omega) \in A\}$. Also we use the notations

$$\begin{aligned}\mathbb{P}(X \in A) &= \mathbb{P}(X^{-1}(A)) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in A\}) \\ \mathbb{P}(X = x) &= \mathbb{P}(X^{-1}(x)) = \mathbb{P}(\{\omega \in \Omega : X(\omega) = x\}).\end{aligned}$$

Notice that we use X to denote the random variable and x to denote its realization (a particular value of X).

Example 2.3. Let X be the number of heads in two fair coin tosses. Then we have

$$\begin{aligned}\mathbb{P}(X = 0) &= \mathbb{P}(\{\omega \in \Omega : X(\omega) = 0\}) = \mathbb{P}(\{TT\}) = 1/4 \\ \mathbb{P}(X = 1) &= \mathbb{P}(\{\omega \in \Omega : X(\omega) = 1\}) = \mathbb{P}(\{TH, HT\}) = 1/2 \\ \mathbb{P}(X = 2) &= \mathbb{P}(\{\omega \in \Omega : X(\omega) = 2\}) = \mathbb{P}(\{HH\}) = 1/4.\end{aligned}$$

Definition 2.4. Given a random variable X, the **cumulative distribution function** (cdf) is the function $F_X : \mathbb{R} \mapsto [0, 1]$ defined by

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

Theorem 2.5. *Let X have a cdf F and let Y have cdf G. If $F(x) = G(x)$ for all x then $\mathbb{P}(X \in A) = \mathbb{P}(Y \in A)$ for all A[5].*

The above theorem says that cdf completely determines the distribution of a random variable.

Theorem 2.6. *(Properties of CDF) A function $F : \mathbb{R} \mapsto [0, 1]$ is a cdf for some probability $\mathbb{P}$ if and only if F satisfies the following three conditions:*
(i) F is non-decreasing: $x_1 < x_2$ implies $F(x_1) \leq F(x_2)$.
(ii) F is normalized:

$$\begin{aligned}\lim_{x \to -\infty} F(x) &= 0 \\ \lim_{x \to \infty} F(x) &= 1.\end{aligned}$$

(iii) F is right-continuous: $F(x) = F(x^+)$ for all x, where $F(x^+) = \lim_{y \downarrow x} F(y)$.

Proof. Omitted. □

Definition 2.7. A random variable X is **discrete** if it takes countably many values. The **probability mass function** (pmf) is then defined as $f_X(x) = \mathbb{P}(X = x)$. We often use $f(x)$ to denote $f_X(x)$ for simplicity.

Example 2.8. Flip a fair coin twice and X be the sum of the heads. Then its pmf is

$$f_X(x) = \begin{cases} 1/4 & x = 0 \\ 1/2 & x = 1 \\ 1/4 & x = 2 \\ 0 & \text{otherwise.} \end{cases}$$

Definition 2.9. A random variable X is continuous if there exists a function f_X such that $f_X(x) \geq 0$ for all x, $\int_{-\infty}^{\infty} f_X(x)dx = 1$ and for every $a \leq b$,

$$\mathbb{P}(a < X < b) = \int_a^b f_X(x)dx.$$

The function f_X is called the probability density function (pdf). Furthermore, we have

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt$$

and $f_X(x) = F_X'(x)$ at all points x at which F_X is differentiable.

Note. For continuous random variable X we have $\mathbb{P}(X = x) = 0$ for every x! Also in the case of continuous variables, $f_X(x)$ does not mean $\mathbb{P}(X = x)$. Actually $f_X(x)$ can take any positive value or even unbounded.

[5]Technically it only holds for every measurable set A.

Example 2.10. Suppose X has pdf

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

Clearly we have $f_X(x) \ge 0$ and $\int f_X(x)dx = 1$. This random variable is said to have a $\mathsf{Uniform}(0,1)$ distribution.

Definition 2.11. Let X be a random variable with cdf F. The **inverse cdf** or **quantile function** is defined by

$$F^{-1}(q) = \inf\{x : F(x) > q\}$$

for $q \in [0,1]$. If F is strictly increasing and continuous then $F^{-1}(q)$ is the unique real number x such that $F(x) = q$. In particular, we call $F^{-1}(1/4)$ the **first quantile**, $F^{-1}(1/2)$ the **median** (or **second quantile**), and $F^{-1}(3/4)$ the **third quantile**.

We use $X \sim F$ to denote that a random variable X has distribution F. In the following we review some important random variables that will be used in this course.

2.2. Some Important Discrete Random Variables.

2.2.1. *The Point Mass Distribution* $X \sim \delta_c$. X has a point mass distribution at a if $\mathbb{P}(X = c) = 1$. Its pmf is

$$f(x) = \begin{cases} 1 & x = c \\ 0 & \text{otherwise.} \end{cases}$$

2.2.2. *The Discrete Uniform Distribution* $X \sim \mathsf{Uniform}(\{c_1, \ldots, c_k\})$. X has a uniform distribution on $\{c_1, \ldots, c_k\}$ if its pmf is given by

$$f(x) = \begin{cases} 1/k & \text{for } x = c_1, \ldots, c_k \\ 0 & \text{otherwise.} \end{cases}$$

2.2.3. *The Bernoulli Distribution* $X \sim \mathsf{Bernoulli}(p)$. X is a Bernoulli random variable with parameter $p \in [0,1]$ if its pmf is given by

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

Sometimes we use the simplified notation $f(x) = p^x(1-p)^{1-x}$ for $x = 0, 1$. Bernoulli random variables are often used to model binary outputs, such as the result of tossing a coin.

2.2.4. *The Binomial Distribution* $X \sim \mathsf{Binomial}(n,p)$. X is a Binomial random variable with parameters $n \in \mathbb{N}$ and $p \in [0,1]$ if its pmf is given by

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

where $\binom{n}{x} = \frac{n!}{x!(n-x)!}$. The Binomial random variable counts the number of successes in n independent Bernoulli random variables with parameter p. Verify that $\sum_{x=0}^{n} f(x) = 1$.

2.2.5. *The Geometric Distribution* $X \sim \mathsf{Geometric}(p)$. X has a Geometric distribution with parameter $p \in (0,1)$ if its pmf is given by

$$f(x) = \begin{cases} p(1-p)^{x-1} & \text{for } x = 1, 2, \ldots \\ 0 & \text{otherwise.} \end{cases}$$

Think of X as the number of flips needed to see a head when flipping a coin. Verify that $\sum_{x=1}^{\infty} f(x) = 1$.

2.2.6. *The Poisson Distribution* $X \sim \mathsf{Poisson}(\lambda)$. X has a Poisson distribution with parameter $\lambda > 0$ if its pmf is given by

$$f(x) = \frac{e^{\lambda}\lambda^x}{x!} \quad x \geq 0.$$

Note that $\sum_{x=1}^{\infty} f(x) = e^{\lambda}\sum_{x=1}^{\infty}\frac{\lambda^x}{x!} = 1$ by the definition of the exponential function. The Poisson is often used to model the counts of rare event.

2.3. **Some Important Continuous Random Variables.**

2.3.1. *The Uniform Distribution* $X \sim \mathsf{Uniform}(a, b)$. X has a $\mathsf{Uniform}(a, b)$ distribution ($a < b$) if its pdf is given by

$$f(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & \text{otherwise.} \end{cases}$$

Its cdf is

$$F(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & x \in [a, b] \\ 1 & x > b. \end{cases}$$

2.3.2. *The Exponential Distribution* $X \sim \mathsf{Exp}(\beta)$. X has an exponential distribution with parameter $\beta > 0$ if its pdf is given by

$$f(x) = \frac{1}{\beta}e^{-x/\beta}, \quad x > 0.$$

It is often used to model the waiting time and has the so-called memoryless property: given $X \sim \mathsf{Exp}(\beta)$ we have $\mathbb{P}(X > t + s|X > t) = \mathbb{P}(X > s)$.

2.3.3. *The Normal/Gaussian Distribution* $X \sim \mathsf{N}(\mu, \sigma^2)$. X has a normal (or Gaussian) distribution with parameters $\mu \in \mathbb{R}$ and $\sigma > 0$ if it has the following pdf

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

Parameter μ is the mean of the distribution and σ is the standard deviation of the distribution (refer to later part of the notes if you do not remember the definitions of mean and standard deviation). X is said to have a **standard normal distribution** if $X \sim \mathsf{N}(0, 1)$. The pdf and cdf of standard normal are denoted by $\phi(z)$ and $\Phi(z)$, respectively.

The normal distribution is the most important distribution in statistics, as many statistics have approximately normal distributions. Below we list some properties of the normal distribution.

Theorem 2.12. *(Properties of the Normal Distribution)*
(1) If $X \sim \mathsf{N}(\mu, \sigma^2)$, *then* $Z = (X - \mu)/\sigma \sim \mathsf{N}(0, 1)$.
(2) If $Z \sim \mathsf{N}(0, 1)$, *then* $X = \mu + \sigma Z \sim \mathsf{N}(\mu, \sigma^2)$.
(3) If $X_i \sim \mathsf{N}(\mu_i, \sigma_i^2)$, $i = 1, \ldots, n$ *are independent, then* $\sum_{i=1}^n X_i \sim \mathsf{N}(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2)$.

3. Bivariate and Multivariate Random Variables

3.1. **Bivariate/Multivariate Distributions.**

Definition 3.1. For any random variables X and Y, the joint distribution function $F(x, y)$ is given by

$$F(x, y) = \mathbb{P}(X \leq x, Y \leq y), \quad x, y \in \mathbb{R}.$$

Similar to the case of univariate random variable, a bivariate (or multivariate) random variable can be discrete, continuous, or neither.

Definition 3.2. Given a pair of discrete random variables X and Y. The **joint probability mass function** for X and Y is given by

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

We often use $f(x, y)$ to denote $f_{X,Y}(x, y)$ for simplicity.

Example 3.3. Flip a unfair coin twice, which has probability 1/3 to be head and 2/3 to be tail. Let X and Y be the results of the first and second flip. Let use 0 to denote "tail" and 1 to denote "head". The joint pmf of (X, Y) is listed in the following table:

	$Y=0$	$Y=1$	
$X=0$	4/9	2/9	2/3
$X=1$	2/9	1/9	1/3
	2/3	1/3	1

Definition 3.4. Let X and Y be continuous random variables with joint distribution function $F(x, y)$. We call a function $f(x, y)$ a **joint pdf** for the random variables (X, Y) if

$$F(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(t_1,t_2)dt_2dt_1, \;\; x,y \in \mathbb{R}.$$

Example 3.5. Let (X, Y) be uniform on the unit square, that is,

$$f_{X,Y}(x,y) = \begin{cases} 1 & x,y \in [0,1] \\ 0 & \text{otherwise.} \end{cases}$$

Clearly we have $\int\int f_{X,Y}(x, y)dxdy = 1$.

All above definitions can be easily generalized to multivariate random variables. For example, the probability distribution function for random variables $\vec{X} = (X_1, X_2, \ldots, X_n)$ is given by

$$F(x_1,\ldots,x_n) = \mathbb{P}(X_1 \le x_1, \ldots, X_n \le x_n), \;\; x_1,\ldots,x_n \in \mathbb{R}.$$

3.2. **Marginal Distributions.**

Definition 3.6. If discrete random variables (X, Y) has joint distribution with pmf $f_{X,Y}(x, y)$, then the **marginal mass functions** for X and $Y$are defined by

$$\begin{aligned} f_X(x) &= \mathbb{P}(X=x) = \sum_y \mathbb{P}(X=x, Y=y) = \sum_y f(x,y) \\ f_Y(y) &= \mathbb{P}(Y=y) = \sum_x \mathbb{P}(X=x, Y=y) = \sum_x f(x,y). \end{aligned}$$

Example 3.7. Suppose $f_{X,Y}$ is given in the table below. Then the marginal mass function for X is the sum of the columns

$$f_X(x) = \begin{cases} 2/3 & x=0 \\ 1/3 & x=1 \\ 0 & \text{otherwise.} \end{cases}$$

and the marginal mass function for Y is the sum of the rows

$$f_Y(y) = \begin{cases} 2/3 & y=0 \\ 1/3 & y=1 \\ 0 & \text{otherwise.} \end{cases}$$

	$Y=0$	$Y=1$	
$X=0$	4/9	2/9	2/3
$X=1$	2/9	1/9	1/3
	2/3	1/3	1

Definition 3.8. If continuous random variables (X, Y) has joint distribution with pdf $f_{X,Y}(x, y)$, then the **marginal density functions** for X and Y are defined by

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

Example 3.9. Let $f_{X,Y}(x, y) = e^{-(x+y)}$ for $x, y \ge 0$. Then we have $f_X(x) = \int_0^\infty f_{X,Y}(x, y)dy = \int_0^\infty e^{-x}e^{-y}dy = e^{-x}$.

3.3. Conditional Distributions.

For discrete random variables X and Y, we already introduced conditional probability $\mathbb{P}(X = x|Y = y)$. Similarly we can define conditional distribution.

Definition 3.10. The **conditional probability mass function** is given by

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

if $f_Y(y) > 0$.

Example 3.11. Toss a coin twice. Let X be the first result (again we use 1 for head, 0 for tail) and Y be the sum of the two results. Then $\mathbb{P}(X = 0|Y = 0) = 1$, $\mathbb{P}(X = 1|Y = 0) = 0$, $\mathbb{P}(X = 0|Y = 1) = 1/2$ and $\mathbb{P}(X = 1|Y = 1) = 1/2$.

The conditional distribution in the continuous case need to be defined in terms of pdf to avoid some technicalities.

Definition 3.12. For continuous random variables X and Y, the **conditional probability density function** is given by

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

assuming that $f_Y(y) > 0$. Then $\mathbb{P}(X \in A|Y = y) = \int_A f_{X|Y}(x|y)dx$.

Example 3.13. Let $X \sim \mathsf{Uniform}(0, 1)$, and given $X = x$, we generate $Y|X = x \sim \mathsf{Uniform}(x, 1)$. Compute the marginal distribution of Y.
Since $f_X(x) = 1$ for $x \in [0, 1]$ and $f_{Y|X}(y|x) = 1/(1 - x)$ for $y \in [x, 1]$, we have

$$f_{X,Y}(x, y) = f_{Y|X}(y|x)f_X(x) = \begin{cases} 1/(1-x) & 0 < x < y < 1 \\ 0 & \text{otherwise.} \end{cases}$$

So the marginal distribution of Y is $f_Y(y) = \int_0^y 1/(1 - x)dx = -\int_1^{1-y} \frac{dt}{t} = -\log(1 - y)$.

3.4. Independent Random Variables.

Definition 3.14. The random variables $X_1, \ldots, X_n$ are **independent** if for all $A_1, \ldots, A_n$ we have

$$\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \mathbb{P}(X_1 \in A_1) \ldots \mathbb{P}(X_n \in A_n).$$

It is difficult to apply the above definition to check independence. Instead, we often use the following theorem.

Theorem 3.15. *Random variables $X_1, \ldots, X_n$ are independent if and only if the cdf can be factorized as*

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

They are also independent if and only if the pdf can be factorized as

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

3.5. Two Important Multivariate Distributions.

3.5.1. *The Multinomial Distribution.* The multinomial distribution is a natural generalization of the binomial distribution.

Definition 3.16. The random vector $\vec{X} = (X_1, \ldots, X_k)$ is said to have a **multinomial** distribution with parameters $n \in \mathbb{N}$ and $p_1, \ldots, p_k$ (where $p_i \geq 0$ for all i and $\sum_{i=1}^k p_i = 1$) if its pmf is given by

$$f(\vec{x}) = \begin{cases} \frac{n!}{x_1! \ldots x_k!} p_1^{x_1} \ldots p_k^{x_k} & \text{if } x_1, \ldots, x_k \in \mathbb{N} \text{ and } \sum x_j = n \\ 0 & \text{otherwise.} \end{cases}$$

Consider drawing a ball from an urn which has balls with k different colors. Let $\vec{p} = (p_1, \ldots, p_k)$ where $p_j \geq 0$ and $\sum_{j=1}^{k} p_j = 1$ and suppose p_j is the probability of drawing a ball of color j. If we draw n times (with replacement) and let $\vec{X} = (X_1, \ldots, X_k)$ where X_j is the number of times that we see a color j ball. Then we say $\vec{X} \sim \mathsf{Multinomial}(n, \vec{p})$.

The multinomial distribution is the multivariate generalization of the binomial distribution (e.g., it specializes to binomial if $k = 2$, $p_1 = p$ and $p_2 = 1 - p$).

Theorem 3.17. *Suppose $\vec{X} \sim \mathsf{Multinomial}(n, \vec{p})$, where $\vec{X} = (X_1, \ldots, X_k)$ and $\vec{p} = (p_1, \ldots, p_k)$. Then the marginal distribution of X_j is $\mathsf{Binomial}(n, p_j)$.*

3.5.2. *The Multivariate Normal Distribution* $\vec{X} \sim \mathsf{N}(\vec{\mu}, \Sigma)$. The random vector $\vec{X} = (X_1, \ldots, X_n)$ has a multivariate normal distribution with parameters $\vec{\mu} \in \mathbb{R}^n$ and Σ (which is a $n \times n$ symmetric, positive definite matrix[11]) has the pdf

$$f_{\vec{X}}(\vec{x}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\vec{x} - \vec{\mu})^T \Sigma^{-1} (\vec{x} - \vec{\mu})\right)$$

where $|.|$ is the matrix determinant and Σ^{-1} is the inverse matrix of Σ.

Similar to the case of a univariate normal random variable, we have

$$\begin{aligned} \mathbb{E}(\vec{X}) &= \vec{\mu} \\ \mathbb{V}(\vec{X}) &= \Sigma. \end{aligned}$$

In particular, $\Sigma_{i,i} = \mathbb{V}(X_i)$ and $\Sigma_{i,j} = \mathsf{Cov}(X_i, X_j)$.

We already know that if two random variables $X_1, \ldots, X_n$ are independent, then $\mathsf{Cov}(X_i, X_j) = 0$ for $i \neq j$. The reverse is not true in general! But if we also know that $\vec{X} = (X_1, \ldots, X_n)$ follows a multivariate normal distribution $\mathsf{N}(\vec{\mu}, \Sigma)$, then the reverse holds.

Theorem 3.18. *If $\vec{X} = (X_1, \ldots, X_n) \sim \mathsf{N}(\vec{\mu}, \Sigma)$ where $\Sigma_{i,j} = 0$ for all $i \neq j$ (e.g., Σ is a diagonal matrix), then $X_1, \ldots, X_n$ are independent.*

It then follows that when Σ is a diagonal matrix with $\Sigma_{i,i} = \sigma_i^2$, we have

$$f_{\vec{X}}(\vec{x}) = \prod_{i=1}^{n} f_{X_i}(x_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(x_i - \mu_i)^2}{2\sigma_i^2}\right).$$

4. Functions of Random Variables

Given a random variable X, let $Y = g(X)$ be a function of X, such as $Y = X^2$. The resulting function Y is also a random variable. The question is, how do we calculate the distribution (pdf/pmf and cdf) of Y?

For the discrete case it can be easily seen that

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

Example 4.1. Let X be the number of heads in two coin tosses. Then we have $f_X(0) = 1/4$, $f_X(1) = 1/2$ and $f_X(2) = 1/4$. If $Y = (X - 1)^2$ then we have $f_Y(0) = 1/2$ and $f_Y(1) = 1/2$.

For continuous case we following three steps to obtain f_Y:

---

1. *For each y, find the set $A_y = \{x : g(x) \leq y\}$.*
2. *Find the cdf by definition*

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \leq y) = \mathbb{P}(\{x : g(x) \leq y\}) = \int_{A_y} f_X(x)dx.$$

[11] This is the only place we use "positive definite" in this course, and there are a few places we use "matrix determinant". Please refer to any linear algebra book for the detailed definitions.

3. Differentiate to get pdf: $f_Y(y) = F'_Y(y)$.

Example 4.2. Let $X \sim \mathsf{Uniform}(-1, 3)$ and find the pdf of $Y = X^2$. The cdf $F_Y(y) = \mathbb{P}(X^2 \le y)$ is easy to compute in separate steps. Clearly $y \in (0, 9)$, and we consider two cases. When $0 < y < 1$ we have $F_Y(y) = \mathbb{P}(-\sqrt{y} \le X \le \sqrt{y}) = \sqrt{y}/2$. When $1 < y < 9$ we have $F_Y(y) = \mathbb{P}(-1 \le X \le \sqrt{y}) = (1 + \sqrt{y})/4$. Take derivative with respect to y we get

$$f_Y(y) = \begin{cases} \frac{1}{4\sqrt{y}} & 0 < y < 1 \\ \frac{1}{8\sqrt{y}} & 1 < y < 9 \\ 0 & \text{otherwise.} \end{cases}$$

The above procedure is applicable to every case. When the function $g(.)$ satisfies certain conditions the calculation can be simplified by the result of the following theorem.

Theorem 4.3. *Let X have pdf $f_X(x)$ and $Y = g(X)$, where g is a strictly monotone increasing or decreasing function. Suppose the inverse g^{-1} is differentiable on the range of X, then the pdf of Y is given by*

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|.$$

Proof. Suppose g is a strictly monotone increasing function, we have

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

and thus

$$f_Y(y) = F'_y(y) = \frac{d}{dy} F_X(g^{-1}(y)) \overset{\text{chain rule}}{=} f_X(g^{-1}(y)) \frac{d}{dy} g^{-1}(y).$$

Similarly we can show that if g is a strictly monotone decreasing function we have $f_Y(y) = -f_X(g^{-1}(y)) \frac{d}{dy} g^{-1}(y)$. By using the property that the derivative of an increasing (decreasing) function g^{-1} is positive (negative) and putting them together we have

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|.$$

□

Example 4.4. Let $f_X(x) = e^{-x}$ for $x > 0$ and let $Y = g(X) = \log X$. Because g is strictly monotone increasing, we have $f_Y(y) = f_X(e^y)e^y = e^y e^{-e^y}$ for $y \in \mathbb{R}$.

We can also apply the above results to functions of random vector (several random variables). For example, if X and Y are random variables, we might want to know $X + Y$, XY, $\max\{X, Y\}$ or $\min\{X, Y\}$. The *three steps procedure* still applies with slight modification:

---

1. For each z, find the set $A_z = \{(x_1, \ldots, x_n) : g(x_1, \ldots, x_n) \le z\}$.

2. Find the cdf by definition

$$\begin{aligned} F_Z(z) &= \mathbb{P}(Z \le z) = \mathbb{P}(g(X_1, \ldots, X_n) \le z) \\ &= \mathbb{P}(\{(x_1, \ldots, x_n) : g(x_1, \ldots, x_n) \le z\}) = \int \cdots \int_{A_z} f_{X_1, \ldots, X_n}(x_1, \ldots, x_n) dx_1 \ldots dx_n. \end{aligned}$$

3. Differentiate to get pdf: $f_Z(z) = F'_Z(z)$.

There is also a multivariate version of theorem 4.3:

Theorem 4.5. *Let $\vec{X} = (X_1, \ldots, X_n)$ be a random vector with pdf $f_{\vec{X}}(x_1, \ldots, x_n)$. Let $\vec{g}(\vec{x}) = (g_1(\vec{x}), \ldots, g_n(\vec{x}))$ where $\vec{g} : \mathbb{R}^n \mapsto \mathbb{R}^n$ is an invertible and differentiable mapping in the range of $\vec{X}$ (one-to-one mapping) then there exists an inverse $\vec{g}^{-1} = (h_1(\vec{y}), \ldots, h_n(\vec{y})) : \mathbb{R}^n \mapsto \mathbb{R}^n$. Let $\vec{Y} = (Y_1, \ldots, Y_n) = \vec{g}(\vec{X})$, then*

$$f_{\vec{Y}}(\vec{y}) = f_{\vec{X}}(\vec{g}^{-1}(\vec{y}))|J|$$

where J is the Jacobian of the inverse mapping defined as

$$J = \begin{vmatrix} \frac{\partial h_1(\vec{y})}{\partial y_1} & \frac{\partial h_1(\vec{y})}{\partial y_2} & \cdots & \frac{\partial h_1(\vec{y})}{\partial y_n} \\ \frac{\partial h_2(\vec{y})}{\partial y_1} & \frac{\partial h_2(\vec{y})}{\partial y_2} & \cdots & \frac{\partial h_2(\vec{y})}{\partial y_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial h_n(\vec{y})}{\partial y_1} & \frac{\partial h_n(\vec{y})}{\partial y_2} & \cdots & \frac{\partial h_n(\vec{y})}{\partial y_n} \end{vmatrix}$$

e.g., J is the determinant of a $n \times n$ matrix.

Proof. Similarly to the univariate case by applying the chain rule. □

Note. There is one inconvenience of the theorem. For example, in order to compute the pdf of $Z = X+Y$ you need to define another random variable such as $W = X - Y$, to use the Jacobian (which is the determinant of a $n \times n$ square matrix).

5. Expectation

5.1. Expectation and Variance.

Definition 5.1. The **expectation** (**mean** or **first moment**) of a random variable X is defined to be

$$\mathbb{E}(X) = \begin{cases} \sum_x x f(x) & \text{if } X \text{ is discrete} \\ \int x f(x) dx & \text{if } X \text{ is continuous.} \end{cases}$$

Note. We assume that the sum or integral exists (well-defined). The expectation is a one-number summary of the distribution that tells the mean or average of a random variable.

Example 5.2. For $X \sim \mathsf{Bernoulli}(p)$ we have $\mathbb{E}(X) = \sum_{x=0,1} x f(x) = 0 \times (1-p) + 1 \times p = p$. For $X \sim \mathsf{Uniform}(a, b)$ we have $\mathbb{E}(X) = \int x f(x) dx = \int_a^b \frac{x}{b-a} dx = \frac{a+b}{2}$.

Theorem 5.3. *(The Law of the Unconscious Statistician) Let $Y = g(X)$. Then the expected value of Y is*

$$\mathbb{E}(Y) = \mathbb{E}(g(X)) = \begin{cases} \sum_x g(x) f(x) & \text{if } X \text{ is discrete} \\ \int g(x) f(x) dx & \text{if } X \text{ is continuous.} \end{cases}$$

Example 5.4. Let $X \sim \mathsf{Uniform}(0, 1)$ and let $Y = g(X) = e^X$. Then we have

$$\mathbb{E}(Y) = \mathbb{E}(e^X) = \int_0^1 e^x dx = e - 1.$$

Alternatively we can first calculate $f_Y(y) = 1/y$ and then $\mathbb{E}(Y) = \int_1^e y \frac{1}{y} dy = e - 1$.

Theorem 5.5. *If $X_1, \ldots, X_n$ are random variables and $c_1, \ldots, c_n$ are constants, then*

$$\mathbb{E}\left(\sum_{i=1}^n c_i X_i\right) = \sum_{i=1}^n c_i \mathbb{E}(X_i).$$

Example 5.6. Let $X \sim \mathsf{Binomial}(n, p)$. Since $X = X_1 + \ldots + X_n$ where $X_i \sim \mathsf{Bernoulli}(p)$ (why?), applying the above rule we have $\mathbb{E}(X) = \sum_{i=1}^n \mathbb{E}(X_i) = np$. Use the pmf of the binomial distribution to verify the result.

Theorem 5.7. *If $X_1, \ldots, X_n$ are **independent** random variables, then*

$$\mathbb{E}\left(\prod_{i=1}^n X_i\right) = \prod_{i=1}^n \mathbb{E}(X_i).$$

Definition 5.8. The **variance** of a random variable X is defined by

$$\mathbb{V}(X) = \mathbb{E}((X - \mathbb{E}(X))^2).$$

The **standard deviation** is $\mathsf{sd}(X) = \sqrt{\mathbb{V}(X)}$.

The variance of a random variable summarizes the scale of the distribution, or how values are spread around the expectation.

Theorem 5.9. *Assuming the variance is well-defined. Then*
(1) $\mathbb{V}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$.
(2) If $X_1, \ldots, X_n$ *are independent random variables and* $c_1, \ldots, c_n$ *are constants, then* $\mathbb{V}(\sum_{i=1}^n c_i X_i) = \sum_{i=1}^n c_i^2 \mathbb{V}(X_i)$.

Note. Unlike the expectation, the summation rule requires the independence. Also notice that $\mathbb{V}(aX + b) = a^2\mathbb{V}(X)$ by treating constant b as a random variable with mean $\mathbb{E}(b) = b$ and $\mathbb{V}(b) = 0$.

Example 5.10. Let $X \sim \mathsf{Binomial}(n, p)$. We have $X = \sum_{i=1}^n X_i$ where X_i's are independent Bernoulli random variables: $X_i \sim \mathsf{Bernoulli}(p)$. So we have $\mathbb{V}(X) = \mathbb{V}(\sum_i X_i) = \sum_i \mathbb{V}(X_i) = \sum_i (\mathbb{E}(X_i^2) - \mathbb{E}(X_i)^2) = np(1-p)$. Use direct calculation to verify the result.

Definition 5.11. The **covariance** of two random variables X and Y is defined as

$$\mathsf{Cov}(X, Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))$$

and the **correlation coefficient** is defined as

$$\rho_{X,Y} = \rho(X, Y) = \frac{\mathsf{Cov}(X, Y)}{\sqrt{\mathbb{V}(X)}\sqrt{\mathbb{V}(Y)}}.$$

Clearly covariance is a generalization of variance, e.g. $\mathsf{Cov}(X, X) = \mathbb{V}(X)$.

Theorem 5.12. *(Properties of Covariance)*
(1) $\mathsf{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$.
(2) If X *and* Y *are independent, then* $\mathsf{Cov}(X, Y) = 0$.
(3) The correlation coefficient satisfies: $-1 \leq \rho_{X,Y} \leq 1$, *and* $|\rho_{X,Y}| = 1$ *if there is a linear relationship between* X *and* Y, *e.g.* $Y = aX + b$.

Note that although independent random variables have covariance 0, but the reverse is not true!

Theorem 5.13. *For random variables* $X_1, \ldots, X_n$,

$$\mathbb{V}\left(\sum_{i=1}^n c_i X_i\right) = \sum_{i=1}^n c_i^2 \mathbb{V}(X_i) + 2\sum_{i=1}^n \sum_{j \neq i} c_i c_j \mathsf{Cov}(X_i, X_j).$$

The concepts of expectation and variance can be easily generalized to random vectors:

Definition 5.14. The **expectation of a random vector** $\vec{X} = (X_1, \ldots, X_n)$ is just the vector of the expectations of each element: $\mathbb{E}(\vec{X}) = (\mathbb{E}(X_1), \ldots, \mathbb{E}(X_n))$. The **variance-covariance matrix** Σ is defined as

$$\Sigma = \mathbb{V}(\vec{X}) = \begin{bmatrix} \mathbb{V}(X_1) & \mathsf{Cov}(X_1, X_2) & \ldots & \mathsf{Cov}(X_1, X_n) \\ \mathsf{Cov}(X_2, X_1) & \mathbb{V}(X_2) & \ldots & \mathsf{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathsf{Cov}(X_n, X_1) & \mathsf{Cov}(X_n, X_2) & \ldots & \mathbb{V}(X_n) \end{bmatrix}.$$

Notice that Σ is symmetric as $\mathsf{Cov}(X_i, X_j) = \mathsf{Cov}(X_j, X_i)$.

5.2. Conditional Expectation and Variance.

Definition 5.15. The conditional expectation of X given $Y = y$ is defined by

$$\mathbb{E}(X|Y=y) = \begin{cases} \sum_x x f_{X|Y}(x|y) & \text{discrete case} \\ \int x f_{X|Y}(x|y)dx & \text{continuous case.} \end{cases}$$

Furthermore, if $g(x)$ is a function of x then

$$\mathbb{E}(g(X)|Y=y) = \begin{cases} \sum_x g(x) f_{X|Y}(x|y) & \text{discrete case} \\ \int g(x) f_{X|Y}(x|y)dx & \text{continuous case.} \end{cases}$$

Note. $\mathbb{E}(X|Y=y)$ is a function of y and $\mathbb{E}(X|Y)$ is a function of the random variable Y. As a result, we have $\mathbb{E}(X|Y)$ itself a random variable. So we can study its mean and variance, etc.

Theorem 5.16. *(The Rule of Iterated Expectation or Double Expectation) For random variables X and Y, assuming the expectations exist, we have*

$$\mathbb{E}(\mathbb{E}(X|Y)) = \mathbb{E}(X).$$

Example 5.17. Let $X \sim \mathsf{Uniform}(0,1)$ and $Y|X = x \sim \mathsf{Uniform}(0,x)$. We have $\mathbb{E}(Y) = \mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}(X/2) = 1/4$. Verify the result by first deriving $f_Y(y)$.

Definition 5.18. The **conditional variance** is defined as

$$\mathbb{V}(X|Y=y) = \begin{cases} \sum_x (x-\mu(y))^2 f_{X|Y}(x|y) & \text{discrete case} \\ \int (x-\mu(y))^2 f_{X|Y}(x|y)dx & \text{continuous case.} \end{cases}$$

where $\mu(y) = \mathbb{E}(X|Y=y)$.

Theorem 5.19. *(Conditional Variance) For random variables X and Y,*

$$\mathbb{V}(X) = \mathbb{E}(\mathbb{V}(X|Y)) + \mathbb{V}(\mathbb{E}(X|Y)).$$

Example 5.20. Let $X \sim \mathsf{Uniform}(0,1)$ and $Y|X = x \sim \mathsf{Uniform}(0,x)$. Compute $\mathbb{V}(Y)$.
We have

$$\begin{aligned} \mathbb{V}(Y) &= \mathbb{E}(\mathbb{V}(Y|X)) + \mathbb{V}(\mathbb{E}(Y|X)) \\ &= \mathbb{E}(X^2/12) + \mathbb{V}(X/2) \\ &= \frac{1}{12} \times \frac{1}{3} + \frac{1}{4} \times \frac{1}{12} \\ &= 7/144. \end{aligned}$$

Verify the result by direct calculation using $f_Y(y)$.

5.3. Moment Generating Functions.

Definition 5.21. The **k-th moment** of a random variable X is defined to be $\mathbb{E}(X^k)$; the **k-th central moment** is defined to be $\mathbb{E}((X - \mathbb{E}(X))^k)$.

It is easy to see that expectation is the 1st moment and variance is the 2nd central moment.

Definition 5.22. The **moment generating function** (mgf) or **Laplace transform** of random variable X is defined by

$$m(t) = \mathbb{E}(e^{tX}), \ \ t \in \mathbb{R}.$$

We say that a moment-generating function for Y exists if there exists an open interval $(-\epsilon, \epsilon)$ such that $m(t)$ is finite for $t \in (-\epsilon, \epsilon)$. In what follows we assume that the mgf exists. The name "moment generating function" comes from the fact that

$$m'(0) = \frac{d}{dt}\mathbb{E}(e^{tX}) \mid_{t=0} = \mathbb{E}\left(\frac{d}{dt}e^{tX}\right) \mid_{t=0} = \mathbb{E}(X).$$

Continue in this way we will get $m^{(k)}(0) = \mathbb{E}(X^k)$, $k = 0, 1, \ldots$.

Theorem 5.23. *(Properties of MGF)*
(1) If $Y = aX + b$, then $m_Y(t) = e^{bt} m_X(at)$.
(2) If $X_1, \ldots, X_n$ are independent and $Y = \sum_{i=1}^n X_i$, then $m_Y(t) = \prod_{i=1}^n m_{X_i}(t)$.

Example 5.24. Let $X \sim \mathsf{Binomial}(n,p)$. Since $X = \sum_{i=1}^{n} X_i$ where $X_i \sim \mathsf{Bernoulli}(p)$ $(i = 1,\ldots,n)$ are independent Bernoulli random variables, we have

$$m_X(t) = \prod_{i=1}^{n} m_{X_i}(t) = (p \times e^t + (1-p))^n.$$

The result of next example is often useful.

Example 5.25. Let $X \sim \mathsf{N}(\mu,\sigma^2)$, then we have

$$\begin{aligned} m_X(t) &= \int_{-\infty}^{\infty} \exp(tx) \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-(\mu+t\sigma^2))^2}{2\sigma^2}\right) \exp\left(\frac{t^2\sigma^2+2\mu t}{2}\right) dx \\ &= \exp\left(\mu t + \sigma^2 t^2/2\right). \end{aligned}$$

Theorem 5.26. *Let X and Y be random variables. If $m_X(t) = m_Y(t)$ for all t in an open interval around 0, then X and Y have the same distribution function (and pdf/pmf).*

The above theorem provides another way to calculate the probability distribution functions of random variables based on the mgf.

Theorem 5.27. *Let $X_i \sim \mathsf{N}(\mu_i,\sigma_i^2)$ be independent random variables for $i = 1,\ldots,n$ and $c_1,\ldots,c_n \in \mathbb{R}$. Then we have*

$$Y = \sum_{i=1}^{n} c_i X_i \sim \mathsf{N}\left(\sum_{i=1}^{n} c_i\mu_i, \sum_{i=1}^{n} c_i^2\sigma_i^2\right).$$

Proof. Let $Y_i = c_i X_i \sim \mathsf{Normal}(c_i\mu_i, c_i^2\sigma_i^2)$. So we have $m_{Y_i}(t) = \exp(c_i\mu_i t + c_i^2\sigma_i^2 t^2/2)$. By the properties of mgf we have

$$m_Y(t) = \prod_{i=1}^{n} m_{Y_i}(t) = \exp\left(\sum_{i=1}^{n} c_i\mu_i t + \sum_{i=1}^{n} c_i^2\sigma_i^2 t^2/2\right)$$

which is the same as the mgf of $\mathsf{Normal}(\sum_i c_i\mu_i, \sum_i c_i^2\sigma_i^2)$ (We used the fact that Y_i's are independent. Why?). By the theorem we conclude that $Y = \sum_i c_i X_i \sim \mathsf{Normal}(\sum_i c_i\mu_i, \sum_i c_i^2\sigma_i^2)$. □