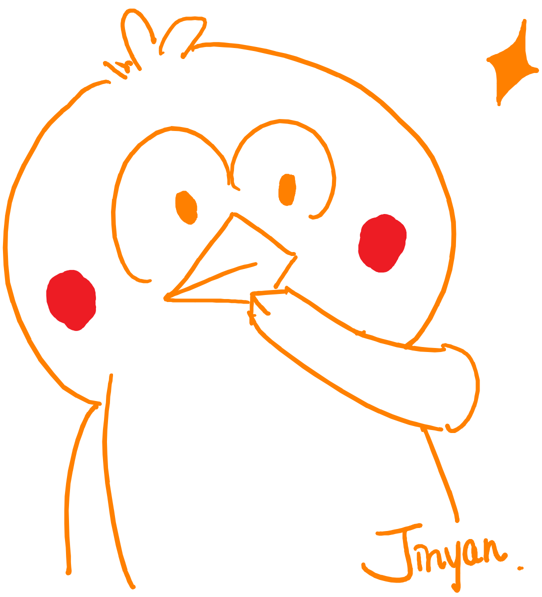
Class Notes

Math 525 - Probability Theory
Professor Dan Burns



Jinyan Miao

Fall 2023

Contents

1 | Probability

There are two basic notions in probability, (1) empirical or sample probability, and (2) models. For (1), one conducts experiments many times and probability is computed by, for instance, number of heads over number of flips of a fair coin. For experiments, one would get a distribution. While on the other hand, (2) is for calculations, but relies on constructions from statistics obtained from experiments.

The Probability Space

Definition 1.0.1
For given an experiment, a set Ω can be used to denote the sample space, which is the collection of all possible outcomes of the experiment. An event E is a subset of Ω.

Example: For flipping a fair coin, one has sample space $\Omega = \{H, T\}$, where H denotes getting a *head* and T denotes getting a *tail*.

Example: For a *height experiment*, one can have

$$\Omega = \{\text{all individuals in the U.S. with their current height}\}.$$

One can say an event $E = \{\text{people with height} \geq 160\,\text{cm}\} \subseteq \Omega$.

Definition 1.0.2
Let Ω be a set, $\mathscr{F}$ is called a σ-algebra on Ω provided that it satisfies:

1. $\emptyset \in \mathscr{F}$
2. *If* $A_n \in \mathscr{F}$ *for all* $n \in \mathbb{N}$, *then* $\bigcup_{n\in\mathbb{N}} A_n \in \mathscr{F}$
3. *If* $A \in \mathscr{F}$, *then* $A^c \in \mathscr{F}$

Remark: Note here from Definition 1.0.2, we have $\emptyset, \Omega \in \mathscr{F}$.

Corollary 1.0.1
Let Ω be a set, and let $\mathscr{F}$ be a σ-algebra on Ω.
If $A_i \in \mathscr{F}$ for all $i \in \mathbb{N}$, then $\bigcap_{i\in\mathbb{N}} A_i \in \mathscr{F}$.
If $A, B \in \mathscr{F}$, then $A \setminus (A \cap B) = A \cap B^c \in \mathscr{F}$.

Definition 1.0.3
Given a sample space Ω, let $\mathscr{F}$ be a σ-algebra of Ω, a probability function $\mathbb{P} : \mathscr{F} \to [0, 1]$ satisfies the conditions

1. $\mathbb{P}(\emptyset) = 0$ *and* $\mathbb{P}(\Omega) = 1$,
2. *If* $A_i \subseteq \Omega$ *are pairwise disjoint for* $i \in \mathbb{N}$, *then* $\mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

Example: Now for an experiment one independently flips two fair coins, then denoting getting a head as H and getting a tail as T, we have $\Omega = \{(T,T),(H,T),(T,H),(H,H)\}$. As the coins are fair, all outcomes in Ω are equally likely. That is each outcome in Ω has probability $1/|\Omega|$. For $E \subseteq \Omega$, one has $\mathbb{P}(E) = |E|/|\Omega|$ where $|\cdot|$ is the cardinality function.

Remark: Probability can be interpreted as an *area* of some subspace of a space. For instance, if $\mathscr{F}$ is a sigma algebra of $\Omega = [0,1]$, then $\mathscr{F}$ contains all intervals $[a,b] \subseteq [0,1]$. Here one could define a probability function by $\mathbb{P}([a,b]) = b - a$, from which we have $\mathbb{P}(\{a\}) = 0$ for $a \in [0,1]$. Note here we have

$$\mathbb{P}\left(\bigcup_{x\in[0,1/2]} \{x\}\right) = \mathbb{P}([0,1/2]) = \frac{1}{2} \neq \sum_{x\in[0,1/2]} \mathbb{P}(\{x\}) = 0$$

holds as the union and sum are not countable.

Definition 1.0.4
Given a sample space Ω, a σ-algebra $\mathscr{F}$ on Ω, and a probability function $\mathbb{P}$ defined on $\mathscr{F}$, the triple $(\Omega, \mathscr{F}, \mathbb{P})$ is called a probability space.

Note: Given a probability space $(\Omega, \mathscr{F}, \mathbb{P})$, if $A, B \in \mathscr{F}$ are disjoint, then it is not hard to see that $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$. While if $A, B \in \mathscr{F}$ have non-empty intersection, then we must write

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)\,,$$

as from the fact that we have

$$A \cup B = (A \setminus (A \cap B)) \cup (B \setminus (A \cap B)) \cup (A \cap B)\,.$$

Lemma 1.0.1
Given a probability space $(\Omega, \mathscr{F}, \mathbb{P})$, for $A, B \in \mathscr{F}$, we obtain

$$\mathbb{P}(A \setminus (A \cap B)) = \mathbb{P}(A) - \mathbb{P}(A \cap B)\,.$$

Conditional Probability

Given a probability space $(\Omega, \mathscr{F}, \mathbb{P})$, note intuitively that one gets immediately

$$\mathbb{P}(A) \cdot \mathbb{P}(B|A) = \mathbb{P}(A \cap B)$$

holds for $A, B \in \mathscr{F}$, where $\mathbb{P}(B|A)$ denotes the probability of B *happening given that* A *has already happened.*

Definition 2.0.1
Given a probability space $(\Omega, \mathscr{F}, \mathbb{P})$, the conditional probability $\mathbb{P}(B|A)$, the probability of $B \in \mathscr{F}$ happening given that $\mathscr{A} \in \mathscr{F}$ has already happened, is formally defined as

$$\mathbb{P}(B|A) := \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)}\,,$$

for which we require that $\mathbb{P}(A) \neq 0$.

Remark: Given a probability space $(\Omega, \mathscr{F}, \mathbb{P})$, if $\mathbb{P}(B) = |B|/|\Omega|$ is satisfied for arbitrary $B \in \mathscr{F}$, one has

$$\mathbb{P}(B|A) = \frac{|A \cap B|/|\Omega|}{|A|/|\Omega|} = \frac{|A \cap B|}{|A|}$$

holds for $A, B \in \mathscr{F}$.

Independence

Definition 3.0.1
Given a probability space $(\Omega, \mathscr{F}, \mathbb{P})$, events $A, B \in \mathscr{F}$ are said to be independent of each other provided that we have $\mathbb{P}(A) \cdot \mathbb{P}(B) = \mathbb{P}(A \cap B)$.

Note: From Definition 2.0.1 and 2.0.2, we see that $A, B \in \mathscr{F}$ are independent, if and only if $\mathbb{P}(B|A) = \mathbb{P}(B)$, if and only if $\mathbb{P}(A|B) = \mathbb{P}(A)$.

Note: Given a probability space $(\Omega, \mathscr{F}, \mathbb{P})$, for $A, B \in \mathscr{F}$, if $A \cap B = \emptyset$, then it is not hard to see that A and B are not independent.

Example: A *fair coin*, formally, is a coin that gives the probability of getting a head as $\mathbb{P}(H) = 1/2$ and the probability of getting a tail as $\mathbb{P}(T) = 1/2$. For two *independent* flips, say flip 1 F_1 and flip 2 F_2, we have

$$\mathbb{P}(F_1 = H,\ F_2 = T) = \mathbb{P}(F_1 = H) \cdot \mathbb{P}(F_2 = T)\,.$$

Example: Now suppose one has a two stage experiment, rolling a single die two times. If the rollings are independent, then the probability of each outcome, say rolling 1 the first time and rolling 5 the second time, denoted as $(1, 5)$, is given by

$$\mathbb{P}((1,5)) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}\,.$$

Example: For another experiment, one rolls a die and denotes the result as x, then rolls a second die, denotes the result as y, and discards that result and roll the second die again if $y > x$. The two rollings here are not independent as $\mathbb{P}((1, 5)) = 0 \neq 1/36$.

Lemma 3.0.1 (Calculating by Conditioning)
Given a probability space $(\Omega, \mathscr{F}, \mathbb{P})$, let $\{E_i \in \mathscr{F} \mid i \in \mathbb{N}\}$ be a collection of events such that E_i are pairwise disjoint and $\bigcup_{i \in \mathbb{N}} E_i = \Omega$. Then the we have

$$\mathbb{P}(A) = \sum_{i \in \mathbb{N}} \mathbb{P}(A|E_i) \cdot P(E_i)\,.$$

Proof. Note here we can write

$$\mathbb{P}(A|E_i) = \frac{\mathbb{P}(A \cap E_i)}{\mathbb{P}(E_i)}\,,$$

hence we have

$$\mathbb{P}(A|E_i) \cdot \mathbb{P}(E_i) = \mathbb{P}(A \cap E_i)\,.$$

Then computing

$$\sum_{i\in\mathbb{N}} \mathbb{P}(A|E_i)\mathbb{P}(E_i) = \sum_{i\in\mathbb{N}} \mathbb{P}(A\cap E_i) = \mathbb{P}\left(\bigcup_{i\in\mathbb{N}} A\cap E_i\right) = \mathbb{P}\left(A\cap\bigcup_{i\in\mathbb{N}} E_i\right) = \mathbb{P}(A)\,,$$

the result then follows. □

Example: Now suppose one has a biased coin that flipping a head has probability $\mathbb{P}(H) = p$ and a tail has probability $\mathbb{P}(T) = 1 - p = q$. Now one flips the coin twice independently, and gets, for instance, a tail first then a head, $\mathbb{P}(H|T) = p \cdot q$.

Canonical samples for experiments. (1) Coin flips. Independent flips of identical coins. (2A) Urn samples with replacement. Given an urn with finite number n of balls, there will be k species of balls, and one samples the urn, makes a draw of 1 ball, puts it back, and makes another draw. (2B) Urn samples, but without replacement, that is drawing one ball, without putting it back, and draw the second ball. Here (1) and (2A) are *independent* experiments, and (2B) is not independent in the two draws.

Now we consider further that in case (2B), we have $k = 2$ labeled as B and R, and $n = 7$, with 3 B's and 4 R's in the urn. Drawing twice without replacement, getting a B in the second draw has the probability

$$\mathbb{P}(B_2) = \mathbb{P}(B_2|R_1)\cdot\mathbb{P}(R_1) + \mathbb{P}(B_2|B_1)\mathbb{P}(B_1) = \frac{1}{2}\cdot\frac{1}{2} + \frac{1}{3}\cdot\frac{3}{7}\,.$$

The probability of getting B at the third drawing without replacement, R in the second drawing without replacement, and R in the first drawing, is given by

$$\mathbb{P}(B_3, R_2, R_1) = \mathbb{P}(B_3|(R_2, R_1))\cdot\mathbb{P}(R_2|R_1)\cdot\mathbb{P}(R_1)\,.$$

Example: Now consider we have an experiment with a fair coin, flipping independently until we get the first head. Let A denotes the arrival flip of getting the first head. Here we have

$$\mathbb{P}(A = 1) = \frac{1}{2}\,, \qquad \mathbb{P}(A = 2) = \frac{1}{4}\,, \qquad \mathbb{P}(A = n) = \frac{1}{2^n}\,.$$

Note that, mathematically,

$$\mathbb{P}(A = \infty) = 1 - \left(\sum_{i\in\mathbb{N}} \mathbb{P}(A = i)\right) = 0\,.$$

Example: Now suppose that we have an n independent flips of a fair coin. The sample space for such an experiment has a size given by

$$|\Omega| = 2^n$$

as we have 2 outcomes, head or tail, for each flip, and a total of n flips. Now for some $0 \le k \le n$, one would like to find the probability of getting k heads in n flips. It is not hard to see here

$$\mathbb{P}(\#(H) = k \text{ in } n \text{ flips}) = \binom{n}{k}\left(\frac{1}{2}\right)^k\left(\frac{1}{2}\right)^{n-k}\,.$$

Note here we have

$$\sum_{k=0}^{k\leq n} \binom{n}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{n-k} = 1 \,.$$

If the coin is instead bias, with probability of getting a head as p and probability of getting a tail as $1-p=q$. The probability becomes

$$\mathbb{P}(\#(H)=k \text{ in } n \text{ flips}) = \binom{n}{k} p^k q^{n-k} \,.$$

One can check that we have

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

These results of flipping a coin tell us how to compute probability for an urn with two colors, draw made with replacement.

Example: Now consider that we have an urn that contains balls of two colors, R and B, and there are r many R's and b and B's in the urn, with $r+b=n$. One makes m draws without replacement, and one would like to find the probability of obtaining k R's and $m-k$ B's. The number of outcomes here, for a total of m draws from an urn with n balls, is $\binom{n}{m}$, and thus we can write

$$\mathbb{P}(\#(R)=k,\ \#(B)=m-k \text{ in } m \text{ draws}) = \frac{\binom{r}{k}\binom{b}{n-k}}{\binom{n}{m}} \,.$$

Example: Now suppose we have a compound experiment. There are three coins, coin 1, denoted as c_1, is fair, coin 2 c_2 is biased with $p=2/3$, and coin 3 c_3 is biased with $p=1/3$. The protocol of the experiment is as follows, first (first step) we flip c_1, if one gets a head then (second step) flip c_2 twice, and if one gets a tail when flipping c_1 (first step), then (second step) flip c_3 twice instead. Let H_1 denote the event that one gets a head in the first flip in the second step. Let H_2 denote the event that one gets a head in the second flip in the second step. Here by conditioning we can write

$$\begin{aligned}\mathbb{P}(H_1) &= \mathbb{P}(H_1|c_1=H)\cdot\mathbb{P}(c_1=H) + \mathbb{P}(H_1|c_1=T)\cdot\mathbb{P}(c_1=T)\\ &= \frac{2}{3}\cdot\frac{1}{2} + \frac{1}{3}\cdot\frac{1}{2} = \frac{1}{2}\,,\end{aligned}$$

$$\begin{aligned}\mathbb{P}(H_2) &= \mathbb{P}(H_2|c_1=H)\cdot\mathbb{P}(c_1=H) + \mathbb{P}(H_2|c_1=T)\cdot\mathbb{P}(c_1=T)\\ &= \frac{2}{3}\cdot\frac{1}{2} + \frac{1}{3}\cdot\frac{1}{2} = \frac{1}{2}\,.\end{aligned}$$

Furthermore, the probability that one gets heads for both flips in the second step is given by

$$\begin{aligned}\mathbb{P}(H_1\cap H_2) &= \mathbb{P}(H_1\cap H_2|c_1=H)\cdot\mathbb{P}(c_1=H) + \mathbb{P}(H_1\cap H_2|c_1=T)\cdot\mathbb{P}(c_1=T)\\ &= \left(\frac{2}{3}\right)^2\cdot\frac{1}{2} + \left(\frac{1}{3}\right)^2\frac{1}{2} = \frac{5}{18}\,.\end{aligned}$$

However, notice that we have $\mathbb{P}(H_1)\cdot\mathbb{P}(H_2) = 1/4$, from which we see that $\mathbb{P}(H_1 \cap H_2) > \mathbb{P}(H_1)\cdot\mathbb{P}(H_2)$, suggesting that H_1 and H_2 are not independent of each other. From Bayes Theorem, we can write

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)}{\mathbb{P}(B)}\,\mathbb{P}(A) = \frac{\mathbb{P}(B|A)\cdot\mathbb{P}(A)}{\mathbb{P}(B|A)\cdot\mathbb{P}(A) + \mathbb{P}(B|A^c)\cdot\mathbb{P}(A^c)}\,, \tag{1.1}$$

where $\mathbb{P}(A)$ is called the prior, $\mathbb{P}(A|B)$ is called the postorior, $\mathbb{P}(B|A)$ is called the likelihood, and $\mathbb{P}(B)$ is called the marginal. From (1.1) we can calculate

$$\mathbb{P}(c_1 = H|H_1 \cap H_2) = \frac{\mathbb{P}(H_1 \cap H_2|c_1 = H)\cdot\mathbb{P}(c_1 = H)}{\mathbb{P}(H_1 \cap H_2)} = \frac{4}{5} > \mathbb{P}(c_1 = H)\,.$$

1.3.1 Independence for more than two events

Definition 3.0.2
Given a probability space $(\Omega, \mathscr{F}, \mathbb{P})$, let $\{A_i\} \subseteq \mathscr{F}$ be a collection of $n \geq 2$ distinct events. We say that events in $\{A_i\}$ are independent of each other provided that we have

$$\mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_n) = \mathbb{P}(A_1)\cdot\mathbb{P}(A_2)\cdot\cdots\cdot\mathbb{P}(A_n)\,.$$

Example: Now suppose we flip a fair coin three times, independently. Here we denote the i-th flip as F_i, here consider three events $A_1 = \{F_2 = F_3\}$, $A_2 = \{F_1 = F_2\}$, and $A_3 = \{F_1 = F_2\}$. The three events are independent pairwise, but the three events are not independent of each other as any two of them imply the third.

2 | Discrete Random Variables

Distributions

Random variables are functions in probability.

Definition 4.0.1
Given a probability space $(\Omega, \mathscr{F}, \mathbb{P})$*, a random variable is a function* $X : \Omega \to \mathbb{R}$ *that satisfies* $\{\omega \in \Omega | X(\omega) \leq x\} \in \mathscr{F}$ *for each* $x \in \mathbb{R}$*, and such function* X *is said to be* $\mathscr{F}$*-measurable.*

Example: One flips a fair coin, and one can define X to be the number of heads in such a flip.

Definition 4.0.2
Given a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ *and a random variable* $X : \Omega \to \mathbb{R}$*, the function* $F : \mathbb{R} \to [0, 1] \quad x \mapsto \mathbb{P}(X \leq x)$ *is called the accumulative distribution function.*

A Bernoulli random variable X defined on $\Omega = \{H, T\}$ is of a form defined by $X(H) = 1$ and $X(T) = 0$. The distribution of X satisfies $F(x) = 0$ for $x < 0$, $F(x) = 1 - p$ for $0 \leq x < 1$ where $p = \mathbb{P}(H)$, and $F(x) = 1$ for $x \geq 1$.

Definition 4.0.3
Given a probability space $(\Omega, \mathscr{F}, \mathbb{P})$*. One can define* $I_A : \Omega \to \mathbb{R}$ *to be a indicator random variable for event* $A \in \mathscr{F}$ *by*

$$I_A(\omega) = \begin{cases} 1 & \omega \in A \\ 0 & \omega \in A^c \end{cases} .$$

Note: From Definition 4.0.4, $\mathbb{P}(I_A = 1) = \mathbb{P}(A)$ and $\mathbb{P}(I_A = 0) = \mathbb{P}(A^c)$.
Remark: All indicator random variables are Bernoulli random variables.

Example: Now suppose one is given two events A and B, here $I_A \cdot I_B = I_{A \cap B}$ is the indicator random variable for the event both A and B happen. Note that $I_A + I_B$ is not necessarily an indicator random variable. $I_A + I_B$ is an indicator random variable if $A \cap B = \emptyset$.

Lemma 4.0.1
Given a probability space $(\Omega, \mathscr{F}, \mathbb{P})$*, and given a random variable* X *with its distribution function* F*, we have (1)* $\mathbb{P}(X > x) = 1 - F(x)$ *for any* $x \in \mathbb{R}$*, (2)* $\mathbb{P}(x < X \leq y) = F(y) - F(x)$ *for any* $x, y \in \mathbb{R}$ *with* $x < y$*, and (3)* $\mathbb{P}(X = x) = F(x) - \lim_{y \to x^-} F(y)$*.*

In general there are several types of random variables, (1) the discrete random variables, and (2) the continuous random variables.

Definition 4.0.4
Given a probability space $(\Omega, \mathscr{F}, \mathbb{P})$*, a random variable* $X : \Omega \to \mathscr{R}$ *is said to be continuous if its distribution function can be expressed as*

$$F(x) = \int_{-\infty}^{x} f(u)\, du$$

for some integrable function $f : \mathbb{R} \to [0, \infty)$ *called the probability density function.*

Definition 4.0.5
Given a probability space $(\Omega, \mathscr{F}, \mathbb{P})$*, a random variable* $X : \Omega \to \mathscr{R}$ *is said to be discrete if it only takes values in some countable subset of* $\mathbb{R}$*. The discrete random variable* X *has proability mass function* $f : \mathbb{R} \to [0, 1]$ *given by* $f(x) = \mathbb{P}(X = x)$*.*

Note: The probability density function for a continuous random variable and the probability mass function for a discrete random variable will be referred as the distribution of the variable, usually denoted as f_X, where X is the random variable.

Example: **Binomial Distribution**. For an experiment of n independent flips of a head-tail coin with $p = \mathbb{P}(H)$, we have $\Omega = \{H, T\}^n$. The total number X of heads takes values in the set $\{0, 1, 2, \cdots, n\}$, and thus is a discrete random variable. The probability mass function is given by

$$f(k) = \mathbb{P}(X = k) = \binom{n}{k} p^k q^{n-k}\,,$$

where $0 \leq k \leq n$. Here the random variable X is said to have the binomial distribution with parameters n and p, denoted as Binomial$[n, p]$, or simply Bin(n, p).

Example: **Uniform Distribution**. Here we consider a random variable which has a uniform distribution $f(x)$ on interval $[a, b] \subseteq \mathbb{R}$. Here we can write

$$\mathbb{P}(\alpha \leq x \leq \beta) = \int_{\alpha}^{\beta} f(x)\, dx\,.$$

Note that $f(x) \geq 0$ for all $x \in \mathbb{R}$ and $\int_{\mathbb{R}} f(x)\, dx = 1$ are required here. Since f is assumed to be uniform, then we can write $f(x) = c$ for all $x \in \mathbb{R}$, where $c = 1/(b - a)$ follows from the normalization of probability.

Example: **Exponential distribution**. Given some probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and a random variable $X : \Omega \to [0, \infty)$ that has a probability density function $f_X(x) = ce^{-\lambda x}$, where λ characterizes the decay rate, and c is required by normalization

$$c = \left(\int_0^{\infty} e^{-\lambda x}\, dx \right)^{-1} = \lambda\,.$$

Here X is called an exponential random variable, the exponential distribution with parameter λ is denoted as Exponential$[\lambda]$, or simply Exp$[\lambda]$.

Example: **Poisson Distribution** If a discrete random variable takes values in the set $\mathbb{N} \cup \{0\}$ with a mass function

$$f_X(k) = \frac{\lambda^k}{k!} e^{-\lambda},$$

with $\lambda > 0$, then the variable X is said to be the Poisson distribution with parameter λ, denoted as Poisson$[\lambda]$.

Multivariate Random Variables

Definition 5.0.1
Let X_i be random variables, $\vec{X} = (X_1, X_2, \cdots, X_n)$ is a random vector, or a multivariate random variable. The joint accumulative distribution function of a random vector $\vec{X}$ on the probability space $(\Omega, \mathscr{F}, \mathbb{P})$ is the function $f_{\vec{X}} : \mathbb{R}^n \to [0,1]$ defined by $F_{\vec{X}}(\vec{x}) = \mathbb{P}(\vec{X} \leq \vec{x})$ for $\vec{x} \in \mathbb{R}^n$.

Note: As before, the expression $\vec{X} \leq \vec{x}$ is an abbreviation for the event $\{\omega \in \Omega | \vec{X}(\omega) \leq \vec{x}\}$ for a random vector $\vec{X}$ on the probability space $(\Omega, \mathscr{F}, \mathbb{P})$.

Lemma 5.0.1
The joint accumulative distribution function $F_{X,Y}$ of the random vector (X, Y) satisfies

1. $\lim_{x,y\to-\infty} F_{X,Y}(x,y) = 0$, $\lim_{x,y\to\infty} F_{X,Y}(x,y) = 1$;
2. *If* $(x_1, y_1) \leq (x_2, y_2)$, *then* $F_{X,Y}(x_1, y_1) \leq F_{X,Y}(x_2, y_2)$;
3. $F_{X,Y}$ *is continuous from above, that is* $\lim_{u,v\to 0^+} F_{X,Y}(x+u, y+v) = F_{X,Y}(x,y)$.

Example: One can think of 2 coins flipped as a bivariate random variable (X_1, X_2), with $X_i(H) = 1$ and $X_i(T) = 0$. Note here the domain of X_i is the result of i-th flip. Here H_2, interpreted as the number of H in two flips, is then given by $H_2 = X_1 + X_2$.

Note: If both X_1 and X_2 are discrete random variables, and the events $\{\omega | X_1(\omega) = x\}$ are independent of the events $\{\omega | X_2(\omega) = y\}$ for all $x, y \in \mathbb{R}$, then we have

$$\mathbb{P}(X_1 = a,\ X_2 = b) = \mathbb{P}(X_1 = a) \cdot \mathbb{P}(X_2 = b).$$

For instance, in the example of flipping two coins, the two flips are independent, but the coin do not have to be fair.

Definition 5.0.2
The random variables X and Y on the probability space $(\Omega, \mathscr{F}, \mathbb{P})$ are said to be jointly discrete provided that (X, Y) takes values in some countable subset of $\mathbb{R}^2$. The jontly discrete random variables (X, Y) have joint probability mass function $f_{X,Y} : \mathbb{R}^2 \to [0,1]$ defined by

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

On the other hand, X and Y are said to be jointly continuous provided that their joint distribution can be expressed as

$$F_{X,Y}(x,y) = \int_{u=-\infty}^{x} \int_{v=-\infty}^{y} f_{X,Y}(u,v)\, du\, dv$$

for $x, y \in \mathbb{R}$ and integrable function $f_{X,Y}$ that is called the joint probability density function of the pair (X, Y).

For this chapter, we will focus on the case where X and Y are discrete random variables, and they are jointly discrete.

Let X and Y have joint mass function $f_{X,Y}$, one can compute the marginal mass functions f_X and f_Y from the knowledge of $f_{X,Y}$,

$$f_X(x) = \mathbb{P}(X = x) = \mathbb{P}\left(\bigcup_y \{X = x\} \cap \{Y = y\}\right) = \sum_y \mathbb{P}(X = x,\, Y = y) = \sum_y f_{X,Y}(x, y)\,.$$

One similarly obtains

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

Definition 5.0.3
Discrete variables X and Y are independent provided that the events $\{X = x\}$ and the events $\{Y = y\}$ are independent for all x and y.

Lemma 5.0.2
The discrete random variables X and Y are independent if and only if

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

for all $x, y \in \mathbb{R}$. More generally, X and Y are independent if and only if $f_{X,Y}(x, y)$ can be factorized as the product $g(x) \cdot h(y)$ of a function of x alone and a function of y alone.

Theorem 5.1 (The Law of Averages)
Let $A_1, A_2, \cdots$ be a sequence of independent events having equal probability $\mathbb{P}(A_i) = p$ where $0 < p < 1$. Denote $S_n = \sum_{i=1}^n I_{A_i}$, the sum of indicator functions of A_i. Here S_n is a random variable which counts the number of occurrences of A_i for $1 \le i \le n$. We have $n^{-1}S_n$ converges to p as $n \to \infty$, in the sense that, for all $\epsilon > 0$, we can write

$$\lim_{n\to\infty} \mathbb{P}(p - \epsilon \le n^{-1}S_n \le p + \epsilon) = 1\,.$$

The Expectation Value

Example: Let X be a random variable that denotes the grade on an exam worth 100 points, for a class of m students $s_1, s_2, \cdots, s_m$, that is $X(s_i) = g_i$ is the grade for the s_i student. The average grade for this class is given by

$$\frac{g_1 + g_2 + \cdots + g_m}{m}\,,$$

or in other words, we can write the average as

$$\sum_{g=0}^{100} \frac{\#(\{s_i \mid X(s_i) = g\}) \cdot g}{m}\,.$$

In this case, the expectation value of X, or the mean, average, of X, is given by

$$\mathbb{E}(X) := \sum_{x_i} \mathbb{P}(X = x_i) \cdot x_i\,.$$

Definition 6.0.1
The mean value, or expectation, or expected value of a discrete random variable X with probability mass function f_X is defined by

$$\mathbb{E}(X) = \sum_{x \in \{x \in \mathbb{R} | f_X(x) > 0\}} x \cdot f_X(x) \,.$$

Example: For the Bernoulli indicator example, X is defined on $\Omega = \{H, T\}$ with $X(H) = 1$ and $X(T) = 0$. The distribution of X satisfies $F(x) = 0$ for $x < 0$, $F(x) = 1 - p$ for $0 \leq x < 1$ where $p = \mathbb{P}(H)$, and $F(x) = 1$ for $x \geq 1$. Here we have

$$\mathbb{P}(X = 0) = q \,, \qquad \mathbb{P}(X = 1) = p \,.$$

Hence we can write

$$\mathbb{E}(X) = \mathbb{P}(X = 0) \cdot 0 + \mathbb{P}(X = 1) \cdot 1 = p \,.$$

Example: For the indicator random variable I_A of an event A, we have

$$\mathbb{E}(I_A) = \mathbb{P}(I_A = 0) \cdot 0 + \mathbb{P}(I_A = 1) \cdot 1 = \mathbb{P}(A)$$

Example: For the binomial distribution, that is $X \sim$Binomial$[n, p]$, [1] we can write

$$\begin{aligned}
\mathbb{E}(X) &= \sum_{k=0}^{n} k \binom{n}{k} p^k q^{n-k} \\
&= p \sum_{k>0}^{n} k \binom{n}{p} p^{k-1} q^{n-k} \\
&= pn \sum_{k>0}^{n} \binom{n-1}{k-1} p^{k-1} q^{n-k} \\
&= pn \sum_{l=0}^{n-1} \binom{n-1}{l} p^{l} q^{n-1-l} \\
&= pn
\end{aligned}$$

Definition 6.0.2
Given a random variable X, denoting $\bar{X} = \mathbb{E}(X)$, we can define

$$Var(X) = \mathbb{E}((X - \bar{X})^2)$$

to be the variance of X. Here $(X - \bar{X})^2$ is a function of X only, so is a random variable.

Note: The variance of a variable describes the dispersion about the mean $\mathbb{E}(X)$ of the given variable X.

Example: Suppose we have $X \sim$Bernoulli$[p]$, here we have $\mathbb{E}(X) = p$, and thus

$$\mathbb{E}((X - p)^2) = (0 - p)^2 \cdot q + (1 - p)^2 \cdot p = p^2 q + q^2 p = pq(p + q) = pq \,.$$

Thus the variance of Bernoulli distribution is pq.

[1] If X is discrete and has a mass function f (or density function if X is continuous), we write $X \sim f$.

Note: We write $X \sim Y$ for random variables X and Y, provided that X and Y are identically distributed, denoted as i.d.. If X and Y are independent and and identically distributed, we denote that they are i.i.d..

Example: For $c = \sum_{n=1}^{\infty} 1/n^2 < \infty$, and a random variable that has a probability mass function $p_X(n) = 1/(cn^2)$ for $n \in \mathbb{N}$, we can write

$$\mathbb{E}(X) = \frac{1}{c} \sum_{n=1}^{\infty} n \frac{1}{n^2} = \infty\,.$$

Lemma 6.0.1
For discrete random variable X, we can write

$$\mathbb{E}(f(X)) = \sum_{x_i} f(x_i)\, p_X(x_i)$$

where x_i are possible values of X, and p_X is the probability mass function of X.

Proof. Here we can write

$$p_{f(X)}(y) = \sum_{f(x)=y} p_X(x)\,.$$

Thus we have

$$\mathbb{E}(f(X)) = \sum_{y} y p_{f(X)}(y)$$

where y is any possible value of $f(x)$ for x being possible values of X, thus

$$\mathbb{E}(f(X)) = \sum_{y} y \sum_{f(x)=y} p_X(x) = \sum_{y} f(x)\, p_X(x)\,,$$

which finishes the proof. Note here for X with an infinity number of possible values, we will enforce $\mathbb{E}(|X|) < \infty$. □

Note: Let (a_i) be a sequence, a rearrangement of (a_i) is any reordering of the terms in (a_i). If $\sum_{i=1}^{\infty} |a_i| < \infty$, then any rearrangement has the same summing result. However, if $\sum_{i=1}^{\infty} a_i$ exists, but is not absolute convergent, that is $\sum_{i=1}^{\infty} |a_i| = \infty$, then one can rearrange the terms such that

$$\sum_{j=1}^{\infty} a_j = c$$

for any $c \in \mathbb{R}$.

Theorem 6.1
Let X and Y be discrete random variables, with finite expectations, then $X + Y$ is also a random variable with expectation

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

Proof. From Lemma 5.0.1, in particular, we can write

$$\begin{aligned}\mathbb{E}(X+Y) &= \sum_{ij}(x_i+y_j)p_{X,Y}(x_i,y_j)\\ &= \sum_{ij} x_i p_{X,Y}(x_i,y_j) + \sum_{ij} y_j p_{X,Y}(x_i,y_j)\\ &= \sum_i x_i p_X(x_i) + \sum_j y_j p_Y(y_j)\\ &= \mathbb{E}(X)+\mathbb{E}(Y)\,,\end{aligned}$$

where $p_{X,Y}$ is the joint mass function of (X,Y), p_X and p_Y are the mass function for X and Y, respectively. We thus complete the proof. □

Note: The statement of Theorem 6.1 does not require that X and Y to be independent.

Theorem 6.2
For random variables X_1, X_2, $\cdots$, X_n, we have

$$\mathbb{E}(X_1+X_2+\cdots+X_n) = \sum_{i=1}^{n}\mathbb{E}(X_i)$$

Example: Consider $X=\sum_{i=1}^{n} X_i$, with X_i being i.i.d. as Bernoulli[p]. Then we have

$$\mathbb{E}(X) = \sum_{i=1}^{n}\mathbb{E}(X_i) = pn$$

Theorem 6.3
For random variables X and Y, we have

$$(\mathbb{E}(XY))^2 \leq \mathbb{E}(X^2)\cdot\mathbb{E}(Y^2)\,,$$

with equality holds if and only if $\mathbb{P}(aX=bY)=1$ for some real a and b, at least one of which is zero.

Proof. We assume that $\mathbb{E}(X^2)$ and $\mathbb{E}(Y^2)$ are strictly positive. Let $Z=aX-bY$ for some $a,b\in\mathbb{R}$. Then we can write

$$0 \leq \mathbb{E}(Z^2) = a^2\mathbb{E}(X^2) - 2ab\mathbb{E}(XY) + b^2\mathbb{E}(Y^2)\,,$$

where we notice here the RHS is quadratic in a with at most one real root. Thus its discriminant must be non-positive, for $b\neq 0$, we obtain

$$(\mathbb{E}(XY))^2 - \mathbb{E}(X^2)\cdot\mathbb{E}(Y^2) \leq 0\,.$$

Furthermore, the discriminant is zero if and only if the quadratic has a real root, and that occurs if and only if $\mathbb{E}(Z^2)=0$, thus obtaining the result. □

The Covaraince

Definition 7.0.1
For random variables X and Y, we define the covariance between X and Y as

$$Cov(X,Y) = \mathbb{E}\left((X - \mathbb{E}(X)) \cdot (Y - \mathbb{E}(Y))\right) .$$

Note: Immediately here we have, for random variable X,

$$\mathrm{Cov}(X,X) = \mathrm{Var}(X) .$$

We see here also

$$\mathbb{E}(\lambda X) = \lambda \mathbb{E}(X) ,$$

and thus

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

Lemma 7.0.1
Let X and Y be independent random variables. Then we can write

$$\mathbb{E}(X \cdot Y) = \mathbb{E}(X) \cdot \mathbb{E}(Y)$$

Proof. Here we can write

$$\begin{aligned}\mathbb{E}(X \cdot Y) &= \sum_{ij} x_i y_j p_{X,Y}(x_i, y_j) \\ &= \sum_{ij} x_i y_j p_X(x_i)\, p_Y(y_j) = \left(\sum_i x_i p_X(x_i)\right) \cdot \left(\sum_j y_j p_Y(y_j)\right) = \mathbb{E}(X) \cdot \mathbb{E}(Y) ,\end{aligned}$$

thus completing the proof. □

Corollary 7.0.1
Let X and Y be independent random variables, then we have

$$Cov(X,Y) = 0 .$$

Proof. Here we can write

$$\mathrm{Cov}(X,Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))) = \mathbb{E}(X - \mathbb{E}(X)) \cdot \mathbb{E}(Y - \mathbb{E}(Y)) = 0$$

□

Lemma 7.0.2
For random variables X and Y, we have

$$Var(X+Y) = Var(X) + 2\,Cov(X,Y) + Var(Y)$$

Proof. Here we can write directly

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

thus completing the proof. □

Lemma 7.0.3
For a random variable Y, we have

$$\mathit{Var}(Y) = \mathbb{E}(Y^2) - \mathbb{E}(Y)^2 \,.$$

Proof. The result follows from the linearity of expectation,

$$\begin{aligned}\mathrm{Var}(Y) &= \mathbb{E}((Y - \mathbb{E}(Y))^2) \\ &= \mathbb{E}(Y^2 - 2Y\mathbb{E}(Y) + \mathbb{E}(Y)^2) \\ &= \mathbb{E}(Y^2) - 2\mathbb{E}(Y)\mathbb{E}(Y) + \mathbb{E}(Y)^2 \\ &= \mathbb{E}(Y^2) - \mathbb{E}(Y)^2 \,.\end{aligned}$$

Thus we obtain our result. □

Example: Consider $X = \sum_{i=1}^{n} X_i$, with X_i being i.i.d. as Bernoulli[p].

$$\mathrm{Var}(X) = \sum_i \mathrm{Var}(X_i) = np(1-p) \,.$$

Example: For geometric random variable X which has mass function $p_X(n) = pq^{n-1}$ with $q = 1 - p$, we can write

$$\sum_{n=1}^{\infty} pq^{n-1} = p \sum_{k=0}^{\infty} q^k = \frac{p}{(1-q)} = 1 \,.$$

Note further that we have

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x} \qquad \text{for } \; -1 < x < 1 \tag{*}$$

Here we can compute

$$\mathbb{E}(X) = \sum_{n=1}^{\infty} npq^{n-1} = p \sum_{n=1}^{\infty} nq^{n-1} = \frac{p}{(1-q)^2} = \frac{1}{p} \,, \qquad \mathrm{Var}(X) = \frac{q}{p^2} \,,$$

where one uses the trick of differentiating, once for expectation and twice for variance, equation (*).

Correlation Coefficient

Let X and Y be random variables. Let $\mu_X = \mathbb{E}(X)$ and $\mu_Y = \mathbb{E}(Y)$. We know further that we have

$$\mathrm{Cov}(X, Y) = \mathbb{E}((X - \mu_X)(Y - \mu_Y)) \,.$$

If X and Y are independent, then $\mathrm{Cov}(X, Y) = 0$, as it follows from

$$\mathbb{E}((X - \mu_X)(Y - \mu_Y)) = \mathbb{E}(X - \mu_X)\,\mathbb{E}(Y - \mu_Y) = 0$$

when X and Y are independent. There are also some dot product properties, for instance, $\mathrm{Cov}(X, Y)$ is bilinear in X and Y, that is

$$\begin{aligned}&\mathrm{Cov}((a_1 X + a_2 X), Y) = a_1 \mathrm{Cov}(X, Y) + a_2 \mathrm{Cov}(X, Y) \,, \\ &\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X) \,, \\ &\mathrm{Var}(X) \cdot \mathrm{Var}(Y) = (\mathrm{Cov}(X, Y))^2 \,.\end{aligned}$$

Definition 8.0.1
The correlation between random variables X and Y is defined by

$$Correlation(X, Y) = \frac{Cov(X, Y)}{\sigma_X \, \sigma_Y} = \rho_{X,Y} \, .$$

where σ_X and σ_Y are the standard deviation of X and Y, respectively, defined by the square root of their variances, that is $\sigma_X = \sqrt{Var(X)}$, and $\sigma_Y = \sqrt{Var(Y)}$.

From the Cauchy-Schwarz Inequality, or Theorem 6.3, it is obvious that we have

$$-1 \leq \rho_{X,Y} \leq 1 \, ,$$

or in other words, $\rho_{X,Y}$ measures the $\cos(\theta)$ for any θ between X and Y as vectors.

Note: For random variable X and Y, $\rho_{X,Y} = 0$ implies $\mathrm{Cov}(X, Y) = 0$, that is X and Y are uncorrelated, but such does not imply X and Y being independent.

Lemma 8.0.1
Let X be a discrete random variable, with $Var(X) = 0$, then $X = \mu_x$ has probability 1.

Proof. Here we can write

$$0 = \mathrm{Var}(X) = \sum_x (x - \mu_x)^2 \, p_X(x) \, ,$$

as $p_X(x)$ and $(x - \mu_x)^2$ are both nonnegative, we require that $x = \mu_x$ for all possible values of x, and thus $X = \mu_x$ has probability 1. $\square$

Theorem 8.1
Let X and Y be random variables. If $\rho_{X,Y} = 1$, then we have

$$\frac{X - \mu_X}{\sigma_X} = \frac{Y - \mu_Y}{\sigma_Y} \, .$$

If we have $\rho_{X,Y} = -1$ instead, then we have

$$\frac{X - \mu_X}{\sigma_X} = -\frac{Y - \mu_Y}{\sigma_Y} \, .$$

Conditional Expectation

Definition 9.0.1
Let X and Y be discrete random variables, we can define a conditional probability mass function from the joint mass function $p_{X,Y}(x,y)$,

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}\,, \qquad \textit{where } p_Y(y) = \sum_x p_{X,Y}(x,y)\,.$$

Example: Now we consider $X \sim \text{Geometric}[p]$.
One would like to compute $\mathbb{P}(X = k+n|X \geq n)$, we write

$$\mathbb{P}(X = k+n|X \geq n) = \frac{\mathbb{P}(\{X = k+n\} \cap \{X \geq n\})}{\mathbb{P}(\{X \geq n\})} = \frac{pq^{n+k-1}}{\sum_{l=n}^{\infty} pq^{l-1}} = \frac{pq^{n+k-1}}{q^{n-1}} = pq^{k-1}$$

where

$$\sum_{l=n}^{\infty} pq^{l-1} = pq^{n-1}\sum_{l=0}^{\infty} q^l = pq^{n-1}\frac{1}{1-q} = q^{n-1}$$

For discrete random variable X and Y, consider the conditional expectation $\mathbb{E}(X|Y)$, which is a random variable as it is a function of Y. For instance, consider the experiment of having $Y \in \{0,1\}$, with $Y \sim \text{Bernoulli}[1/2]$ and X being the count of flipping coins. If $Y = 0$, then we flip coin$_0$ which has probability $p = 2/3$ twice independently. If $Y = 1$, then we flip coin$_1$ which has probability of getting a head $p = 1/3$ twice independently. X is the count of head obtained in the two flips of a coin. Here we can write

$$\mathbb{E}(X|Y=0) = 2 \cdot \frac{1}{3} = \frac{2}{3}\,, \qquad \mathbb{E}(X|Y=1) = 2 \cdot \frac{2}{3} = \frac{4}{3}\,.$$

Here we can further compute

$$\begin{aligned}\mathbb{E}(\mathbb{E}(X|Y)) &= \sum_y \mathbb{E}(X|Y=y) \cdot p_Y(y) = \mathbb{E}(X) \\ &= \mathbb{E}(X|Y=0) \cdot p_Y(0) + \mathbb{E}(X|Y=1) \cdot p_Y(1) \\ &= \frac{2}{3} \cdot \frac{1}{2} + \frac{4}{3} \cdot \frac{1}{2} = 1\,.\end{aligned}$$

Thus we see here it is intuitive to view $\mathbb{E}(X|Y)$ as a random variable taking values in the image of Y.

Theorem 9.1
Given discrete random variables X, Y, we have

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

Proof. Here we can write

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

then we have

$$\mathbb{E}(\mathbb{E}(X|Y)) = \sum_y \sum_x x\,\frac{p_{X,Y}(x,y)}{p_Y(y)}\,p_Y(y) = \sum_x \sum_y x\,p_{X,Y}(x,y) = \sum_x x p_x(x) = \mathbb{E}(X)\,.$$

The result follows. □

Theorem 9.2
Given discrete random variables X, Y, we have

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

for any function g for which both expectations exist.

Note that the variance of a binomial distribution

$$\mathrm{Var}\left(\mathrm{Binomial}[n,p]\right) = npq\,.$$

Example: Consider an compound experiment. Here $Y \sim \mathrm{Uniform}[0,1]$, and X is the count of two independent flip of a single coin which has a probability of getting a head as $p = Y$. Here we can write

$$\mathbb{E}(\mathbb{E}(X|Y)) = \int_0^1 \mathbb{E}(X|Y=p)\,dp = \int_0^1 2p\,dp = 1\,.$$

Random Walk

Here we consider a symmetric random walk on a line. Starting at the origin $x=0$, one flips a fair coin. If one gets a head, then one walks to the right $x+1$, and if one gets a tail instead, then one walks to the left $x-1$. One is interested in his position after n flips. Let X_n denote the position after n flips. Here we see that $X_1 \in \{-1,1\}$, with probability $\mathbb{P}(X_1=-1) = 1/2$ and $\mathbb{P}(X_1=1) = 1/2$. Thus here $X_1 = 2Y-1$ where $Y \sim \mathrm{Bernoulli}[1/2]$. Let Y_i be i.i.d. $\sim \mathrm{Bernoulli}[1/2]$, then $2Y_i - 1$ can denote the i-th step in the walk, and thus we can write

$$X_n = \sum_{i=1}^n (2Y_i - 1) = \left(2\sum_{i=1}^n Y_i\right) - n\,.$$

Thus we obtain

$$\mathbb{E}(X_n) = 2\cdot\mathbb{E}\left(\sum_{i=1}^n Y_i\right) - n = 2\cdot\frac{n}{2} - n = n - n = 0\,.$$

Note in fact, there is a theorem saying that, with probability 1, one will be able to visit $|x| \geq \sqrt{n}$ for large enough n. This is an example of a random process, in particular a discrete random process, with discrete $X_1,\ X_2,\ X_2,\ \cdots,\ X_n,\ \cdots$ of random variables, and here n is called the discrete time. The random process describes the probabilistic dynamics, that is the evolution of probability in time n. In our case of random walk, once the initial point is picked, the sample space consists of ordered sequences of heads and tails of the coin, or translated into ordered sequences of steps.

Another way to describe the random walk is by considering Y_i to be i.i.d. $\sim \mathrm{Bernoulli}[p]$,

$$X_i = 2Y_i - 1 = \begin{cases} 1 & \mathbb{P} = p \\ -1 & \mathbb{P} = 1-p = q \end{cases},$$

we then denote

$$S_n = \sum_{j=1}^n X_j = \text{location of the walker on the line after } n \text{ steps.}$$

Assume here $S_0 = a$ being the initial condition, for some $a \in \mathbb{Z}$. Notice that the random walk is spatially homogeneous as we have

$$\mathbb{P}(S_n = b|S_0 = a) = \mathbb{P}(S_n = (b-a)|S_0 = 0)\,.$$

The random walk is also temporally homogeneous, that is we have

$$\mathbb{P}(S_n = b|S_0 = a) = \mathbb{P}(S_{n+k} = b|S_k = a)$$

for some $k \in \mathbb{N} \cup \{0\}$.

Example: Gambler's Ruin problem.
Here we have two players A and B, a referee who flips a fair coin independently. If the referee gets a head, then A wins \$1 from B, and if the referee gets a tail, then B wins \$1 from A. If any one of them wins all money from the other one, then the game ends. The problem, is quite similar to the random walk problem, but with end points at both sides of the x-axis, representing the amount of money that the two persons have initially. Note here $\pm(M+N)$ is called the absorbing barriers, where M is the money that A has initially, and N is the money that B has initially.

Definition 10.0.1
A general definition for a random process is $\{X_n\}$, a set of random variables. The random process is said to be Markov provided that

$$\mathbb{P}(X_n|X_0, X_1, \cdots, X_{n-1}) = \mathbb{P}(X_n|X_{n-1})\,.$$

That is, a random process is said to be Markov provided that the past history is not needed to determine the probability, only the current state. A Markov process is said to be stationary provided that $\mathbb{P}(X_n|X_{n-1})$ is independent of n.

Example: For the random walk example (second approach, using S_n to denote the location of walker on the line after n steps), we have

$$\mathbb{P}(S_n|S_{n-1}) = \mathbb{P}(S_n|S_0, S_1, \cdots, S_{n-1})\,,$$

that is the random walk is a Markov process. Notice further that, if $p = 1/2 = q$, the probability of any path γ from a to a reachable point b in n steps is equal to any other points that is reachable from a in n steps, as we have here

$$\mathbb{P}(\gamma) = p^r q^{n-r} = p^r q^l$$

where $n - r = l$. In the discussion below we impose the probability $p = 1/2$. Here each path γ of the walker is called a sample path. Let $N_n(a, b)$ denote the number of paths from a to b in exactly n steps, let $(0, a)$ denote the initial condition, that is position at a at the 0-th step, and (n, b) denote the terminal position at b at the n-th step. Then we can write

$$N_n(a, b) = \binom{n}{(n+b-a)/2}\,.$$

As one starts at a, the net motion of the walker is $b - a$. Notice here we have $r + l = n$, and $b - a = r - l$, then we can write

$$n + (b-a) = 2r\,,$$

rearranging we obtain

$$r = \frac{1}{2}(n+b-a)\,, \qquad\qquad l = \frac{1}{2}(n-b+a)\,.$$

One problem of interest is whether the walker return to 0 before N. For this, we want to count the number of paths from a to b which visit 0 in between. Notice immediately that if a and b carry opposite signs, then all paths from a to b will cross 0. For $a, b > 0$, we denote the number of paths in n steps from a to b which visit 0 as $N_n^0(a,b)$. One can show that we have

$$N_a^0(a,b) = N_n(-a,b)$$

as any path connecting a and b that visits 0 can be decomposed into two paths: (1) a path connecting a and 0 and (2) a path connecting 0 and b. Simply reflecting path (1) with respect to the spatial axis, with $p = 1/2 = q$, we obtain the desired result. As a corollary, with $b > 0$, the number of paths from 0 to b which do not revisit 0 in n steps is given by

$$N_n(0,b) - N_n^0(0,b) = \frac{b}{n}N_n(a,b)\,,$$

which is always an integer as one can show. To obtain $(b/n)N_n(a,b)$, one must start with the first step going in the positive direction as $b > 0$, the number of such paths is given by $N_{n-1}(1,b) - N_{n-1}^0(1,b)$, that is we can write

$$\begin{aligned} N_{n-1}(1,b) - N_{n-1}^0(1,b) &= \binom{n-1}{(n-1+b-1)/2} - N_{n-1}(-1,b) \\ &= \binom{n-1}{(n+b-2)/2} - \binom{n-1}{(n-1+b+1)/2} = \frac{b}{n}\binom{n}{(n+b)/2}\,. \end{aligned}$$

Assuming that we start from 0, note that we can write

$$\mathbb{P}(S_n = b,\ S_1, S_2, \cdots, S_n \neq 0) = \frac{|b|}{n}N_n(0,b)p^r q^l = \frac{|b|}{n}\mathbb{P}(S_n = b)$$

for $|r - l| = |b|$ and $r + l = n$. As a corollary, we have

$$\mathbb{P}(S_1, S_2, \cdots, S_n \neq 0) = \frac{\mathbb{E}(|S_n|)}{n}\,,$$

as we can write here

$$\begin{aligned} \mathbb{P}(S_1, S_2, \cdots, S_n \neq 0) &= \sum_b \mathbb{P}(S_1, S_2, \cdots, S_n \neq 0, S_n = b) \\ &= \sum_b \frac{|b|}{n}\mathbb{P}(S_n = b) \\ &= \frac{1}{n}\sum_b |b|\mathbb{P}(S_n = b) \\ &= \frac{\mathbb{E}(|S_n|)}{n} \end{aligned}$$

Next we consider the problem where $M_n \coloneqq \max_{1\leq i\leq n} S_i$. We will show here we have

$$\mathbb{P}(M_n \geq r,\, S_a = b) = \begin{cases} \mathbb{P}(S_n = b) & r \leq b \\ (q/p)^{r-b}\mathbb{P}(S_n = 2r-b) & b < r \end{cases}$$

Let $N_n^r(b)$ denote the number of paths c from 0 to b which visit r in between, that is $S_i = r$ for some $0 < i < n$. For any such a path, one can reflect about r the portion of the path that is between the last time it reaches r and the final position at b, then if $b < r$ the probability

$$\begin{aligned} \mathbb{P}(M_n \geq r,\, S_n = b) &= (\text{number of paths } c)\, p^k q^l \\ &= N_n(0, 2r-b) p^{k'} q^{l'} q^{r-b} p^{-r+b} \\ &= (q/p)^{r-b}\mathbb{P}(S_n = 2r-b) \end{aligned}$$

where $k + l = k' + l' = n$, $k' - l' = 2r - b$, and $k - l = b$. Now we consider the problem of hitting time for b, that is the first n such that $S_n = b$. Here we set $F_b(n)$ to be the probability of n being the first arrival at b. Then we can write

$$\begin{aligned} F_b(n) &= \mathbb{P}(M_{n-1} = S_{n-1} = b-1,\ S_n = b) \\ &= p\,(\mathbb{P}(M_{n-1} \geq b-1,\ S_{n-1} = b-1) - \mathbb{P}(M_{n-1} \geq b,\, S_{n-1} = b-1)) \\ &= p\,(\mathbb{P}(S_{n-1} = b-1) - (q/p)\mathbb{P}(S_{n-1} = b+1)) \\ &= p\mathbb{P}(S_{n-1} = b-1) - q\mathbb{P}(S_{n-1} = b+1) \\ &= |b|\mathbb{P}(S = b)/n\,. \end{aligned}$$

3 Continuous Random Variables

Distribution and Expectation

A continuous random variable X is a random variable whose distribution $F_X(x) \coloneqq \mathbb{P}(X \leq x)$ is a continuous function of x. Let $f_X(x) \geq 0$ be the probability density function of x defined on $\mathbb{R}$, then we can write

$$\mathbb{P}(X \leq x) = \int_{-\infty}^{x} f_X(t)\, dt$$

Note here $f_X(x)$ further satisfies

$$\lim_{x \to \infty} \int_{-\infty}^{x} f_X(t)\, dt = 1\,, \qquad \lim_{x \to -\infty} \int_{-\infty}^{x} f_X(t)\, dt = 0\,.$$

Example: For uniform distribution between a and b with $a < b$, denoted as Uniform$[a, b]$, it is easy to verify that the density function reads

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

as required by the normalization of probability. Here we can write

$$F_X(x) = \int_{-\infty}^{x} \frac{dt}{b-a} = \begin{cases} 0 & x \leq a \\ \frac{x-a}{b-a} & a \leq x \leq b \\ 1 & x \geq b \end{cases}.$$

Example: For the exponential distribution whose density function is defined by

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

one can compute that we have

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt = \lambda \int_{0}^{x} e^{-\lambda t}\, dt = 1 - \lambda \int_{x}^{\infty} e^{-\lambda t}\, dt = 1 - e^{-\lambda x}\,,$$

for $x \geq 0$.

Example: We consider normal random variable which has a density function

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

Note here we can write

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-x^2/2} = \frac{2}{\sqrt{2\pi}}\int_{0}^{\infty} e^{-x^2/2}\,dx$$

as $f_X(x)$ is an even function. Furthermore, we have

$$\begin{aligned}\left(\frac{2}{\sqrt{2\pi}}\int_{0}^{\infty} e^{-x^2/2}\,dx\right)^2 &= \frac{2}{\pi}\int_{0}^{\infty}\int_{0}^{\infty} e^{-(x^2+y^2)/2}\,dx\,dy\\ &= \frac{2}{\pi}\int_{0}^{\pi/2}\int_{0}^{\infty} e^{-r^2/2} r\,dr\,d\theta\\ &= \frac{2}{\pi}\left[-e^{-r^2/2}\right]_0^{\infty}\int_{0}^{\pi/2} d\theta = 1\,.\end{aligned}$$

Thus we see that $f_X(x)$ is indeed normalized. Here X is called Z, the standard unit normal. We then can write

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

Definition 11.0.1
For a continuous random variable X, we can similarly define the expectation of X as

$$\mathbb{E}(X) \coloneqq \int_{-\infty}^{\infty} x\, f_X(x)\,dx\,.$$

The variance of X is defined by

$$Var(X) = \int_{-\infty}^{\infty} (x - \mathbb{E}(X))^2\; f_X(x)\,dx\,.$$

Lemma 11.0.1
For discrete random variable X, we have

$$\mathbb{E}(f(X)) = \sum_x f(x)\,p_X(x)\,.$$

For continuous random variable X, we have similarly

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

One can show equivalently that we have

$$\mathrm{Var}(X) = \int_{-\infty}^{\infty} x^2 f_X(x)\,dx - \left(\int_{-\infty}^{\infty} x\,f_X(x)\,dx\right)^2$$

for a continuous random variable X.

Example: Let $X \sim \mathrm{Uniform}[a,b]$, we have $\mathbb{E}(X) = (a+b)/2$, as we can write

$$\mathbb{E}(X) = \frac{1}{b-a}\int_a^b x\,dx = \frac{1}{b-a}\frac{b^2-a^2}{2} = \frac{b+a}{2}\,.$$

The variance can also be computed

$$\begin{aligned}\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X))^2 &= \frac{1}{b-a}\int_a^b x^2\,dx - \left(\frac{a+b}{2}\right)^2 \\ &= \frac{1}{3}\frac{b^3-a^3}{b-a} - \frac{a^2+2ab+b^2}{4} = \frac{(b-a)^2}{12}\,.\end{aligned}$$

Example: Now let $X \sim \text{Exponential}[\lambda]$, we have

$$\mathbb{E}(X) = \int_0^\infty \lambda x e^{-\lambda x}\,dx = \int_0^\infty e^{-\lambda x}\,dx = \frac{1}{\lambda}\,,$$

where we have utilized integration by parts for integrating.

Example: Now consider the standard unit normal Z, we can compute

$$\mathbb{E}(Z) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^\infty x e^{-x^2/2}\,dx = 0$$

by the even property of density function of Z. We can also compute

$$\mathrm{Var}(Z) = \mathbb{E}(Z^2) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^\infty x^2 e^{-x^2/2}\,dx = 1\,.$$

Example: For normal distribution, more generally, we can have

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

where μ is the mean of the distribution, and σ^2 is the variance.

In general, if we have a continuous distribution $X \sim f_X(x)$, and we define a new random variable $(X-a) \sim f_X(x+a)$. Note here $X-a$ has density function $f_X(x+a)$ as we can write

$$F_{X-a}(x) = \mathbb{P}(X - a \le x) = \mathbb{P}(X \le x+a) = \int_{-\infty}^{x+a} f_X(t)\,dt\,,$$

thus we see here

$$f_{X-a}(x) = \frac{d}{dx}F_{X-a}(x) = f_X(x+a)\,.$$

Similarly, we have $f_{X+a}(x) = f_X(x-a)$. It is not hard to see that we have

$$\mathbb{E}(X-a) = \mathbb{E}(X) - \mathbb{E}(a) = \mathbb{E}(X) - a\,,$$

as we can write

$$\int x\,f_X(x+a)\,dx = \int (x-a)\,f_X(x)\,d(x-a) = \int x\,f_X(x)\,dx - a\int f_X(x)\,dx = \mathbb{E}(X) - a\,.$$

If one scales the random variable instead, that is, consider λX for some nonzero positive constant λ, then we see here

$$F_{\lambda X}(x) = \mathbb{P}(\lambda X \le x) = \mathbb{P}(X \le x/\lambda)\,,$$

thus we can write

$$f_{\lambda X}(x) = \frac{d}{dx} F_{\lambda X}(x) = \frac{d}{dx} \int_{-\infty}^{x/\lambda} f_X(x)\, dx = \frac{1}{\lambda} f_X(x/\lambda)\,.$$

It follows here, for the standard unit normal Z, we have

$$f_{\lambda Z}(x) = \frac{1}{\sqrt{2\pi\lambda^2}} e^{-x^2/(2\lambda^2)}$$

which has expectation 0 and variance λ^2. Here we thus write

$$f_{\lambda Z+\mu}(x) = \frac{1}{\sqrt{2\pi\lambda^2}} e^{-(x-\mu)^2/(2\lambda^2)}\,,$$

which has expectation μ and variance λ^2.

Definition 11.0.2
For joint distribution of two continuous variables X and Y, the density function is denoted as $f_{X,Y}(x, y)$, with marginals defined by

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

Definition 11.0.3
Similar to the discrete variables, we can define the conditional density of two continuous variables X and Y,

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

Suppose here we are given jointly distributed random variables X and Y, one might be interested in finding the density for $Z = X + Y$. We first consider the accumulative distribution function $F_Z(z)$ and take $(d/dz)F_Z(z)$ to get the density function $f_Z(z)$. Here we write

$$F_Z(z) = \mathbb{P}(Z \leq z) = \int_{-\infty}^{z} f_Z(\zeta)\, d\zeta\,,$$

thus we have

$$\frac{dF_Z}{dz}|_z = f_Z(z)\,.$$

Here we are remained to find $\mathbb{P}(Z \leq z)$, note that we have

$$\mathbb{P}(Z \leq z) = \mathbb{P}(X + Y \leq z) = \int_{H_z} f_{X,Y}(x, y)\, dx\, dy$$

where $H_z = \{(x, y)|x + y \leq z\} \subseteq \mathbb{R}^2$. Thus we obtain

$$\mathbb{P}(Z \leq z) = \int_{-\infty}^{\infty} \int_{-\infty}^{z-y} f_{X,Y}(x, y)\, dx\, dy\,.$$

Here we consider a change of variable $u = x + y$ and $y = y$, with Jacobian

$$\frac{\partial(x, y)}{\partial(u, y)} = \det\left(\begin{bmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial y} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial y} \end{bmatrix}\right) = \begin{vmatrix} 1 & -1 \\ 0 & 1 \end{vmatrix} = 1\,.$$

Thus we see here

$$\mathbb{P}(Z \leq z) = \int_{-\infty}^{z} \int_{-\infty}^{\infty} f_{X,Y}(u - y, y)\, dy\, du\,,$$

from which we can write

$$f_{X+Y}(z) = \frac{d}{dz}\left(\int_{-\infty}^{z} \int_{-\infty}^{\infty} f_{X,Y}(u - y, y)\, dy\, du\right) = \int_{-\infty}^{\infty} f_{X,Y}(z - y, y)\, dy\,.$$

Note further that, when X and Y are independent, then the joint distribution satisfies

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

and the distribution of $f_{X+Y}(z)$ becomes

$$f_{X+Y}(z) = \underbrace{\int_{-\infty}^{\infty} f_X(z - y)\, f_Y(y)\, dy}_{\text{convolution of } f_X,\, f_Y}\,.$$

If X and Y are discrete variables instead, then the mass function of $X + Y$ is just

$$p_{X+Y}(z) = \sum_{x+y \leq z} p_{X,Y}(z - y, y)\,.$$

Example: Let X, Y be i.i.d. $\sim$ Exponential$[\lambda]$. That is,

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

Then $X + Y$ will be the *waiting time* for two independent variables. Note that the density functions for the individual variables X and Y vanish for values less than zero, thus for $z \geq 0$, we have

$$\begin{aligned} f_{X+Y}(z) = \int_0^{\infty} f_X(z - y) f_Y(y)\, dy &= \int_0^{z} f_X(z - y)\, f_Y(y)\, dy \\ &= \int_0^{z} \lambda^2 e^{-\lambda(z-y)-\lambda y}\, dy \\ &= \lambda^2 e^{-\lambda z} \int_0^{z} dy = \lambda^2 z e^{-\lambda z}\,, \end{aligned}$$

which is an example of Gamma distribution, characterized by

$$f_\Gamma(z) = \begin{cases} \lambda^n \frac{z^{n-1}}{(n-1)!} e^{-\lambda z} & z \geq 0 \\ 0 & z < 0 \end{cases}$$

with $n = 2$.

Example: Consider now U being uniform random variable on $[-1/2, 1/2]$, with $X_j \cdots$ i.i.d. $\sim U$. Let $S_n = \sum_{j=1}^{n} X_j$, we will see later that when n is large enough, the distribution of S_n approaches a normal distribution.

Geometric Probability

Here events are considered to be subsets of $\mathbb{R}^n$. First we consider the example of Buffon's Needle. Consider boards of widths w, positioned side by side with width going in the vertical direction, and infinite long in the horizontal direction, and needle of length L, dropped randomly to the boards. We are interested in the probability of needle touching the crack (edge between two boards). Here we assume that $L < w$. Let x be, from the bottom of a board, of the midpoint of the needle, thus $0 \leq x \leq w$. Let θ be the angle between the needle and the horizontal axis, $0 \leq \theta \leq \pi$. Note here the needle will cross the upper crack of a board if we have

$$w - x < \frac{L}{2} \sin(\theta) \, ,$$

and will cross the lower creak of a board if we have

$$x < \frac{L}{2} \sin(\theta) \, .$$

Note that, by randomness of dropping the needle, x and θ are independent and uniform distributed, over $[0, w]$ and $[0, \pi]$, respectively. Thus here we can write

$$\mathbb{P}(\text{needle on a crak}) = \frac{2}{w\pi} \int_0^{\pi} \frac{L}{2} \sin(\theta) \, d\theta = \frac{2L}{w\pi} \, .$$

Sums of Normals

We denote a normal distribution as $\text{Normal}[\mu, \sigma^2]$, where μ is the mean and σ^2 is the variance of the distribution. Here we consider independent $X_1, X_2 \sim \text{Normal}[0, 1]$. The distribution of $X_1 + X_2$ should be, intuitively, normal, and have greater variance, in fact, $X_1 + X_2 \sim \text{Normal}[0, 2]$. In general, if we have $Y_1 \sim \text{Normal}[0, \sigma_1]$ and $Y_2 \sim \text{Normal}[0, \sigma_2^2]$, then we have in fact

$$Y_1 + Y_2 \sim \text{Normal}[0, \sigma_1^2 + \sigma_2^2] \, .$$

Furthermore, if we have $W_1 \sim \text{Normal}[\mu_1, \sigma_1^2]$, $W_2 \sim \text{Normal}[\mu_2, \sigma_2^2]$, then we have

$$W_1 + W_2 \sim \text{Normal}[\mu_1 + \mu_2, \, \sigma_1^2 + \sigma_2^2] \, , \qquad (*)$$

where we have assumed here Y_i, W_i, and X_i are all independent of each other.

For the proof of the sum of normals, we first consider $X_i \sim \text{Normal}[0, 1]$, then we can write

$$f_{X_i}(x_i) = \frac{1}{\sqrt{2\pi}} e^{-x_i^2/2} \, ,$$

thus the distribution of $f_{X_1+X_2}(z)$ gives

$$\begin{aligned} f_{X_1+X_2}(x) = \int_{-\infty}^{\infty} f_{X_1}(x-y)\, f_{X_2}(y)\, dy &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-(x-y)^2/2} e^{-y^2/2}\, dy \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-((x-y)^2+y^2)/2}\, dy \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-x^2/4-(y-x/2)^2}\, dy \\ &= \frac{1}{2\pi} e^{-x^2/4}\int_{-\infty}^{\infty} e^{-(y-x/2)^2}\, dy \\ &= \frac{1}{2\pi} e^{-x^2/4}\int_{-\infty}^{\infty} e^{-y^2}\, dy \\ &= \frac{1}{\sqrt{4\pi}} e^{-x^2/4} \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-y^2/2}\, dy \\ &= \frac{1}{\sqrt{4\pi}} e^{-x^2/4}\,, \end{aligned}$$

which corresponds to a normal distribution of mean 0 and variance 2, that is $(X_1+X_2) \sim$ Normal$[0,2]$. Similarly, one can show (*) using similar procedure. Furthermore, for $W \sim \text{Normal}[\mu,\sigma^2]$, we see here

$$\begin{aligned} f_{-W}(w) = \frac{d}{dw}\int_{-\infty}^{w} f_{-W}(t)\, dt &= \frac{d}{dw}\mathbb{P}(-W \le w) \\ &= \frac{d}{dw}\mathbb{P}(W \ge -w) = \frac{d}{dw}\int_{-w}^{\infty} f_W(t)\, dt = f_W(-w)\,, \end{aligned}$$

thus we see here

$$(-W) \sim \text{Normal}[-\mu,\sigma^2]\,.$$

Example: Consider $X, (Y-X) \sim \text{Exponential}[\lambda]$, where X is the waiting time for the first arrival, and Y is the waiting time for the second arrival. Note that, as Y is the time for the second arrival, thus we must have $X \le Y$, with the fact that X and $(Y-X)$ are independent of each other, we can write

$$f_Y(z) = \int_0^{\infty} \lambda e^{-\lambda(z-y)}\, \lambda e^{-\lambda y}\, dy = e^{-\lambda z}\lambda^2 \int_0^z dy = z e^{-\lambda z}\lambda^2\,,$$

whenever $z \ge x$. Now note that we have

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

which agrees with our observation that, as $Y-X$ is independent of X,

$$f_{Y|X}(y|x) = \begin{cases} \lambda e^{-\lambda(y-x)} & y \ge x \\ 0 & y < x \end{cases}\,,$$

and thus

$$f_{X,Y}(x,y) = \lambda e^{-\lambda(y-x)}\lambda e^{-\lambda x} = \lambda^2 e^{-\lambda y}\,,$$

whenever $y \geq x$. Finally we obtain

$$f_Y(y) = \int_0^y f_{X,Y}(x,y)\,dx = \int_0^y \lambda^2 e^{-\lambda y}\,dx = \lambda^2 y e^{-\lambda y}\,,$$

whenever $y \geq 0$.

Now we consider independent multivariate normals, with X, Y independently $\sim$ Normal$[0,1]$, then we can write

$$f_{X,Y}(x,y) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} = \frac{1}{2\pi} e^{-(x^2+y^2)/2}\,.$$

Note in general for for X, Y independently distributed as some Normal$[0, \sigma_i^2]$, then we can write their joint distribution

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_1\sigma_2} e^{-(x^2/\sigma_1^2 + y^2/\sigma_2^2)/2}\,.$$

Level-sets of $f_{X,Y}(x,y) = c$ in $\mathbb{R}^2$ can be seen as ellipses

$$\frac{x^2}{\sigma_1^2} + \frac{y^2}{\sigma_2^2} = k\,,$$

for $k = 1$, we see immediately that the ellipse intersects $(-\sigma_1, 0),\ (0, \sigma_2),\ (\sigma_1, 0),\ (0, -\sigma_2)$.

Definition 13.0.1
In general, for X_i distributed as normals $[\mu_i, \sigma_i^2]$, we can write

$$f_{X_1,X_2,\cdots,X_n} = \frac{1}{(2\pi)^{n/2}(\det(A))^{1/2}} e^{-(\vec{x}-\vec{\mu})^T A(\vec{x}-\vec{\mu})/2}\,,$$

where A is a symmetric positive definite $n \times n$ matrix, $A = (a_{ij})$, such that we have

$$(\vec{x}-\vec{\mu})^T A(\vec{x}-\vec{\mu}) = \sum_{ij} (x_i - \mu_i) a_{ij} (x_j - \mu_j)\,.$$

The matrix $V = A^{-1}$ is called the covariance matrix.

Note: If the off diagonal terms of A are non-zero, then the marginals are not independent. For instance, in the case where $n = 2$ and X_1, X_2 both have expectations 0, the level sets of $f_{X_1,X_2} = c$ is characterized completely by

$$\sum_{ij} x_i a_{ij} x_j = k\,,$$

which are ellipses but with axes not necessarily parallel to the coordinate axes. We can in fact find random variables U and V being independent from each other here via change of variables, such that the the level sets of the joint distribution of U and V have ellipse given by $au^2 + bv^2 = k$, with $a, b > 0$. Further, we will see that $U = \alpha X + \beta Y$ and $V = \gamma X + \delta Y$.

Consider here the joint distribution for some symmetry positive definite matrix A,

$$f_{X_1,X_2,\cdots,X_n}(\vec{x}) = \frac{1}{(2\pi)^{n/2}(\det(A))^{1/2}} e^{-\vec{x}^T A \vec{x}/2}\,.$$

Note that a symmetric positive definite matrix A has eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_n > 0$ with corresponding eigenvectors $v_1, v_2, \cdots, v_n$ such that

$$Av_i = \lambda_i v_i \,,$$

with $v_i \cdot v_j = \delta_{ij}$ provided that $\lambda_i \neq \lambda_j$ when $i \neq j$. That is $\{v_i\}$ is an orthonormal basis of $\mathbb{R}^n$. One can construct O from $\{v_i\}$ such that we have

$$O^T O = I \,, \qquad\qquad AO = O\Lambda \,,$$

where I is the identity and Λ is the diagonal matrix with λ_i on the diagonal. Note also that $\det(O) = \pm 1$. Now we see that we have

$$\vec{x}^T A \vec{x} = \vec{x}^T \left(O\Lambda^{1/2}\right) \left(O\Lambda^{1/2}\right)^T \vec{x} = \left((O\Lambda^{1/2})^T \vec{x}\right)^T \cdot \left(O\Lambda^{1/2}\right)^T \vec{x} = \vec{y}^T \vec{y}$$

where we have defined

$$\vec{y} = \left(O\Lambda^{1/2}\right)^T \vec{x} \,, \qquad \Lambda^{1/2} = \mathrm{diag}(\lambda_i^{1/2}) \,.$$

In other words, the change of variable satisfies

$$\vec{x} = O\Lambda^{1/2}\vec{y}$$

Combining, we obtain

$$f_{Y_1,Y_2,\cdots,Y_n}(\vec{y}) = \frac{1}{(2\pi)^{n/2}(\det(A))^{1/2}} e^{-(\lambda_1^2 y_1^2 + \lambda_2^2 y_2^2 + \cdots + \lambda_n^2 y_n^2)/2} \,.$$

For $X_1, X_2, \cdots, X_n$ being i.i.d.. One can define

$$\bar{X} = \frac{1}{n}\sum_n X_i\,,$$

such that $\mathbb{E}(\bar{X})$ serves as an estimation on $\mathbb{E}(X_i)$, as we can write

$$\mathbb{E}(\bar{X}) = \mathbb{E}\left(\frac{1}{n}\sum_n X_i\right) = \frac{n}{n}\,\mathbb{E}(X_i) = \mathbb{E}(X_i)\,,$$

thus $\bar{X}$ is an unbiased estimator of $\mu = \mathbb{E}(X_i)$. Note further here, as X_i are independent, we obtain

$$\mathrm{Var}(\bar{X}) = \frac{1}{n^2}\,\mathrm{Var}\left(\sum X_i\right) = \frac{1}{n^2}\sum_i \mathrm{Var}(X_i) = \frac{1}{n}\,\sigma^2\,.$$

Now consider the definition

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

First we consider the case where $n = 2$, we write

$$S_2^2 = (X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 = \left(\frac{1}{2}(X_1 - X_2)\right)^2 + \left(\frac{1}{2}(X_2 - X_1)\right)^2 = \frac{1}{2}\,(X_1 - X_2)^2\,.$$

Note that, if the mean of X_1 and X_2 are both 0, then we can write

$$\mathbb{E}\left(\frac{1}{2}(X_1 - X_2)^2\right) = \frac{1}{2}\mathbb{E}((X_1 - X_2)^2) = \frac{1}{2}\,(\mathrm{Var}(X_1) + \mathrm{Var}(X_2)) = \mathrm{Var}(X_i)\,,$$

Thus we see here $S_n^2/(n-1)$ gives an estimation for $\mathrm{Var}(X_i)$. In fact, one can check that $S_n^2/(n-1)$ is an unbiased estimator.

Example: Consider $X \sim N(\mu, \sigma^2)$ and X_i being i.i.d. $\sim X$, then it is not hard to verify

$$\bar{X} \sim N\left(\mu,\, \sigma^2/n\right)\,.$$

Further, we can verify that

$$\frac{1}{n-1}S_n^2 \sim \frac{\sigma^2}{(n-1)^2}\,\chi^2(n-1)\,,$$

where $\chi^2(n)$ is the chi-squared with n degrees of freedom, defined by

$$\chi^2(n) = \sum_{i=1}^{n} Y_i^2$$

where Y_i are i.i.d. $\sim Y \sim N(0,1)$. Note that for $n = 2$, $\chi^2(2)$ can be interpreted as characterizing the squared distance from the origin of the random vector (Y_1, Y_2) in $\mathbb{R}^2$.

Here we denote $W \sim \chi^2(1) \sim Z^2$ where $Z \sim N(0,1)$. One would be interested in the density function of $\chi^2(1)$, that is the density function of W. It is obvious that $f_W(x) = 0$ for $x < 0$, for $w \geq 0$, we can write

$$f_W(w) = \frac{d}{dw} F_W(w) = \frac{d}{dw}\, \mathbb{P}(W \leq w)\,,$$

where we have

$$\mathbb{P}(W \leq w) = \mathbb{P}(-\sqrt{w} \leq Z \leq \sqrt{w}) = \frac{1}{\sqrt{2\pi}} \int_{-\sqrt{w}}^{\sqrt{w}} e^{-x^2/2}\, dx\,.$$

Thus combining we obtain

$$f_W(w) = \frac{d}{dw}\left(\frac{2}{\sqrt{2\pi}} \int_0^{\sqrt{w}} e^{-x^2/2}\, dx\right) = \frac{2}{\sqrt{2\pi}} e^{-(\sqrt{w})^2/2}\, \frac{d}{dw}(\sqrt{w}) = \frac{1}{\sqrt{2\pi}} e^{-w/2} w^{-1/2}\,.$$

Example: Here we can look at the exponential distribution $U \sim$ Exponential$[\lambda]$, and U_i being i.i.d. $\sim U$. One can interpret U as modeling a waiting time for the first arrival of a count in a decay experiment. Thus U_1 is the waiting time for the first arrival, $U_1 + U_2$ is that for the second arrival, and $\sum_{i=1}^{n} U_i$ is that of the n-th arrival. Furthermore, U is a continuous analogue of the discrete geometric variable with parameter p, which has an expectation $1/p$, thus it is not hard to see that U has expectation $1/\lambda$. Therefore, $\sum_i U_i$ are discrete analogue of the negative binomial distributions. Here $U_1 + U_2$ has the density function

$$f_{U_1+U_2}(x) = \int_{-\infty}^{\infty} f_{U_1}(x-y)\, f_{U_2}(y)\, dy = \int_0^x \lambda^2 e^{-\lambda(x-y)}\, e^{-\lambda y}\, dy = \lambda^2 x e^{-\lambda x}\,,$$

for $x \geq 0$ and vanishes for $x < 0$. The density function of $U_1 + U_2 + U_3$ is

$$f_{U_1+U_2+U_3}(x) = \int_{-\infty}^{\infty} f_{U_1+U_2}(y)\, f_{U_3}(x-y)\, dy = \int_0^x \lambda^2 y e^{-\lambda y} \lambda e^{-\lambda(x-y)}\, dy = \frac{\lambda^3 x^2}{2} e^{-\lambda x}\,,$$

for $x \geq 0$ and vanishes for $x < 0$. Proceed via induction, one finds that we have

$$f_{U_1+U_2+\cdots+U_n}(x) = \begin{cases} 0 & x < 0 \\ (\lambda^n x^{n-1} e^{-\lambda x})/(n-1)! & x \geq 0 \end{cases},$$

which is a Gamma distribution $\Gamma(n, \lambda)$ with parameters n and λ.

Example: **Gamma Distribution**

For any real number $t > 0$, $W \sim \Gamma(t, \lambda)$ has the density function given by

$$f_W(x) = \frac{\lambda^t x^{t-1} e^{-\lambda x}}{\Gamma(t)}\,,$$

where the Gamma function $\Gamma(t)$ is defined by

$$\Gamma(t) = \int_0^{\infty} s^{t-1} e^{s}\, ds\,.$$

One can show in fact we have $\Gamma(n) = (n-1)!$. First we will show that $\Gamma(t+1) = t\Gamma(t)$.

$$\Gamma(k+1) = \int_0^\infty x^k e^{-x}\,dx = \left[-x^k e^{-x}\right]_0^\infty + \int_0^\infty k t^{k-1} e^{-x}\,dx = k\,\Gamma(k)\,.$$

Thus as a consequence, we have

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

Furthermore, we have

$$\Gamma(1) = \int_0^\infty s^0 e^{-s}\,ds = \int_0^\infty e^{-s}\,ds = 1\,,$$

thus we conclude that we have

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

To match the $\chi^2(1)$ distribution with a Gamma distribution, we observer that we require $t = 1/2$ and $\lambda = 1/2$ for parameters in the Gamma distribution, and we need to verity $\Gamma(1/2) = \sqrt{\pi}$. Here we compute

$$\begin{aligned}
\Gamma(1/2) &= \int_0^\infty s^{-1/2} e^{-s}\,ds \\
&= \int_0^\infty \left(\frac{u^2}{2}\right)^{-1/2} e^{-u^2/2}\,u\,du \\
&= \int_0^\infty \sqrt{2} e^{-u^2/2}\,du \\
&= \frac{\sqrt{2}}{2}\int_{-\infty}^\infty e^{-u^2/2}\,du = \sqrt{\pi}\,.
\end{aligned}$$

where we have utilized change of variable $s = u^2/2$. Thus we have verified that $\chi^2(1) = \Gamma(1/2, 1/2)$. By induction, one can obtain that

$$\chi^2(x) = \Gamma\left(\frac{n}{2}, \frac{1}{2}\right)\,.$$

Here we have further that

$$\Gamma(n+1/2) = (n-1/2)(n-3/2)\cdots(1/2)\sqrt{\pi}\,.$$

Now suppose we have $X \sim N(\mu, \sigma^2)$, and X_i being i.i.d. $\sim X$. As we have previously discussed, we want to show that

$$\bar{X} \sim N(\mu,\, \sigma^2/n)\,, \qquad\qquad \frac{S_n^2}{(n-1)} \sim \frac{\sigma^2}{(n-1)^2}\chi^2(n-1)\,,$$

and that $\bar{X}$ and $S_n^2/(n-1)$ are independent. WLOG, we can assume that $\mu = 0$ and $\sigma^2 = 1$. Otherwise we can replace X_i by $Y_i = (X_i - \mu)/\sigma^2$. As we have previously shown, the joint density function is invariant by rotations. that is, if one takes A to be an orthogonal matrix $A^T A = \mathbb{I}$, then one would like to find a new collection of random variables $Z_1, Z_2, \cdots, Z_n$ such that $Z_i = \sqrt{n}\bar{X} \sim N(0,1)$.

Recall the definition

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

for $X_i \sim N(\mu, \sigma^2)$, is an estimator for μ, and that the definition

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

is an estimation of variance σ^2. WLOG, we can assume here $\mu = 1$ and $\sigma^2 = 1$, then we can replace X_i by

$$Y_i = \frac{X_i - \mu}{\sigma} \sim N(0,1)\,,$$

and it can be shown that $S^2 \sim \chi^2(n-1)$, with

$$S^2 = \sum_{i=2}^{n} Z_i^2\,,$$

with $Z \sim N(0,1)$. Further, one can verify that $\bar{X}$ and $S^2/(n-1)$ are independent.

The strategy is to utilize change of variable. Here we choose

$$Z_1 = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i = \sqrt{n}\bar{Y}\,.$$

That is, if one consider $Y_i := e_i$, we have

$$Z_1 = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} e_i := v_1\,,$$

where we would like to find an orthonormal basis $\{v_1, v_2, \cdots, v_n\}$. From which one can form an orthogonal matrix A defined by

$$A = \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix}\,,$$

here note that $A^T = A^{-1}$. In which case one can write

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = A^{-1}\begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix}\,.$$

A change of random variables from Y to Z gets us that

$$\frac{1}{(\sqrt{2\pi})^n}e^{-|y|^2/2} = \frac{1}{(\sqrt{2\pi})^n}e^{-|A^t z|^2/2}\cdot|\det(A^t)| = \frac{1}{(\sqrt{2\pi})^n}e^{-|z|^2/2}\,,$$

from which case one can conclude

$$\sum_{i=1}^{n} Y_i^2 = \sum_{i=1}^{n} Z^2 = Z_1^2 + \sum_{i=2}^{n} Z_i^2\,.$$

Notice that we have

$$
\begin{aligned}
\sum_{i=2}^{n} Z_i^2 &= \left(\sum_{i=1}^{n} Y_i^2\right) - Z_1^2 \\
&= \left(\sum_{i=1}^{n} Y_i^2\right) - \frac{1}{n}\left(\sum_{i=1}^{n} Y_i\right)^2 \\
&= \left(\sum_{i=1}^{n} Y_i^2\right) - \frac{n^2}{n}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right)^2 \\
&= \left(\sum_{i=1}^{n} Y_i^2\right) - n(\bar{Y})^2 \\
&= \sum_{i=1}^{n}(Y_i^2 - \bar{Y}^2) \\
&= \sum_{i=1}^{n}(Y_i^2 - \bar{Y}^2) - 2\sum_{i=1}^{n} Y_i\bar{Y} + 2\sum_{i=1}^{n} Y_i\bar{Y} \\
&= \sum_{i=1}^{n}(Y_i - 2Y_i\bar{Y} - \bar{Y}^2) + 2\sum_{i=1}^{n} Y_i\bar{Y} \\
&= \sum_{i=1}^{n}(Y_i - 2Y_i\bar{Y} - \bar{Y}^2) + 2n\bar{Y}^2 \\
&= \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \\
&= S^2 \,.
\end{aligned}
$$

4 Limit Theorems

Generating Functions

Generating function is the idea that we take some data and incorporate them into a single function. For instance, suppose X is a discrete random variable with values in $\mathbb{N} \cup \{0\}$, with mass function given by $p_X(i)$. We define the probability generating function

$$G_X(s) = \sum_{i=0}^{\infty} p_X(i) s^i \,,$$

then if X is Bernoulli$[p]$, which takes value 0 and 1, with probability $p_X(0) = q$ and $p_X(1) = p$, we here have

$$G_X(s) = q + ps \,.$$

Definition 1.0.1
We define here the probability generating function

$$G_X(s) = \mathbb{E}(s^X) \,,$$

for a random variable X.

If $Y_1, Y_2, \cdots, Y_n$ are independent random variables, and $Y = \sum_{i=1}^{n} Y_i$, one can show that we have

$$G_Y(s) = \prod_{i=1}^{n} G_{Y_i}(s)$$

To show that is holds for Bernoulli random variables, we proceed by induction, here for the case where $n = 2$. First note that we have

$$G_{Y_1+Y_2}(s) = \sum_{i=0}^{\infty} p_{Y_1+Y_2}(i)\, s^i$$

From the definition of G_{Y_1} and G_{Y_2}, it is not hard to verify, in the case where Y_i are Bernoulli random variables, that we have

$$\begin{aligned} G_{Y_1}(s)\, G_{Y_2}(s) &= p_{Y_1+Y_2}(0) + (p_{Y_1}(0)\, p_{Y_2}(1) + P_{Y_1}(1)\, p_{Y_2}(0))\, s \\ &= p_{Y_1+Y_2}(0) + p_{Y_1+Y_2}(1)s = G_{Y_1+Y_2}(s) \,. \end{aligned}$$

Example: Applied to $X \sim \text{Binomial}[n,p]$, we have that

$$G_X(s) = G_{X_1+X_2+\cdots+X_n}(s) = G_{X_1}(s)\cdot G_{X_2}(s)\cdot \cdots \cdot G_{X_n}(s) = (q+ps)^n\,,$$

where $X_i \sim \text{Bernoulli}[p]$.

Example: Consider the $X \sim \text{Geometric}[p]$, then $p_X(n) = pq^{n-1}$. We can write

$$G_X(s) = \sum_{n=1}^{\infty} pq^{n-1}s^n = ps\sum_{n=1}^{\infty} q^{n-1}s^{n-2} = ps\sum_{j=0}^{\infty}(qs)^j = \frac{ps}{1-qs}\,.$$

Definition 1.0.2
We define here the moment generating function

$$M_X(s) = \sum_{n=0}^{\infty}\frac{s^n}{n!}\,\mathbb{E}(X^n) = \mathbb{E}(e^{sX})\,,$$

for a random variable X.

Here one can also show that we have

$$M_{X+Y}(s) = M_X(s)\cdot M_Y(s)$$

for independent random variables X and Y.

Example: For X ~Bernoulli$[p]$, we have that

$$M_X(t) = \mathbb{E}(e^{tX}) = e^{t\cdot 0}\cdot q + e^t p = q + pe^t\,.$$

Example: For X ~Binomial$[n,p]$, which can be equivalently written as

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

for X_i i.i.d. ~Bernoulli$[p]$, we have

$$M_X(t) = (q+pe^t)^n\,.$$

Example: For X ~Geometric$[p]$, we have that

$$\begin{aligned}
M_X(t) &= \sum_{n=1}^{\infty} e^{tn}pq^{n-1}\\
&= pe^t\sum_{n=1}^{\infty} e^{t(n-1)}q^{n-1}\\
&= pe^t\sum_{j=0}^{\infty} e^{tj}q^j\\
&= pe^t\sum_{j=0}^{\infty}(e^tq)^j = \frac{pe^t}{1-qe^t}\,,
\end{aligned}$$

which does not always converge.

Example: For U ~Uniform$[a, b]$, we have that

$$M_U(t) = \frac{1}{b-a} \int_a^b e^{tx}\, dx = \frac{1}{b-a} \left[\frac{1}{t} e^{tx}\right]_a^b = \frac{1}{b-a} \left(\frac{e^{tb}}{b} - \frac{e^{ta}}{a}\right).$$

Example: For X ~Exponential$[\lambda]$, we have that

$$M_X(t) = \lambda \int_0^\infty e^{tx} e^{-\lambda x}\, dx = \lambda \int_0^\infty e^{-(\lambda-t)x}\, dx = \frac{\lambda}{\lambda - t} \left[-e^{-(\lambda-t)x}\right]_0^\infty ,$$

which exists only when $\lambda - t > 0$, in which case we have

$$M_X(t) = \frac{\lambda}{\lambda - t}.$$

Example: For X ~Normal$[\mu, \sigma^2]$, we have that

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

Here we use a trick

$$\begin{aligned}
-\frac{(x-\mu)^2}{2\sigma^2} + xt &= -\frac{1}{2\sigma^2} \left((x-\mu)62 - 2\sigma^2 tx\right) \\
&= -\frac{1}{2\sigma^2} \left(x^2 - 2\mu x + \mu^2 - 2\sigma^2 + x\right) \\
&= -\frac{1}{2\sigma^2} \left(x^2 - 2(\mu+\sigma^2 t)x + (\mu+\sigma^2 t)^2 + \mu^2 - (\mu+\sigma^2 t)^2\right) \\
&= -\frac{1}{2\sigma^2} \left(x - (\mu+\sigma^2 t)\right)^2 - \frac{1}{2\sigma^2} \left(\mu^2 - \mu^2 - 2\mu\sigma^2 t - \sigma^4 t^2\right) ,
\end{aligned}$$

thus we have

$$\begin{aligned}
M_X(t) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty e^{-(x-(\mu+\sigma^2 t))^2/(2\sigma^2)} e^{(2\mu\sigma^2 t+\sigma^4 t^2)/(2\sigma^2)}\, dx \\
&= e^{\mu t+\sigma^2 t^2/2} \frac{1}{\sqrt{2\pi}} \int e^{-(x-(\mu+\sigma^2 t))^2/(2\sigma^2)}\, dx \\
&= e^{\mu t+(\sigma^2 t^2)/2}.
\end{aligned}$$

Note a key observation here we have

$$\ln\left(M_X(t)\right) = \mu t + \frac{\sigma^2 t^2}{2} ,$$

which does not depend on higher degrees of t.

Generating Function in Random Walk

Consider here the simple random walk, where X_i i.i.d. $\sim X$, with X defined by

$$X = \begin{cases} 1 & \text{with probability } p \\ -1 & \text{with probability } q = 1-p \end{cases}.$$

Denote here $S_n = S_0 + \sum_{i=1}^{n} X_i$, with $S_0 = 0$. The problem of interest is the probability of first return back to zero. That is, we write

$$f_0(n) = \text{probability that walker's first return to zero is at time } n$$

we note immediately that we have $f_0(n) = 0$ for odd $n \in \mathbb{N}$, and we enforce that $f_0(0) = 0$. We also write

$$p_0(n) = \text{proabability that } S_n = 0\,.$$

Thus we have

$$f_0(n) = \mathbb{P}(S_n = 0,\ S_1, S_2, \cdots, S_{n-1} \neq 0)\,.$$

We define here

$$P_0(s) = \sum_{n=0}^{\infty} p_0(n)\, s^n\,, \qquad F_0(s) = \sum_{n=0}^{\infty} f_0(n)\, s^n\,.$$

We would first like to find $p_0(n)$ by conditioning on events $0 \leq k \leq n$ is the first return to 0. That is

$$p_0(n) = \sum_{k=0}^{n} p_0(n-k)\, f_0(k)\,,$$

here we see that

$$s^n p_0(n) = \sum_{k=0}^{n} s^{n-k}\, p_0(n-k)\, s^k f_0(k)\,,$$

from which we obtain

$$P_0(s) = \sum_{n=0}^{\infty} s^n p_0(n) = 1 + \sum_{n=0}^{\infty} \sum_{k=0}^{n} s^{n-k} p_0(n-k)\, s^k\, f_0(k) = 1 + P_0(s) \cdot F_0(s)\,.$$

While we have

$$P_0(s) = \sum_{n=0}^{\infty} p_0(n)\, s^n = \sum_{n=0}^{\infty} p_0(2n)\, s^{2n} = \sum_{n=0}^{\infty} \binom{2n}{n} (pq)^n\, s^{2n} = (1 - 4pqs^2)^{-1/2}\,.$$

Combining all, one finds that we have

$$P_0(s) = (1 - 4pqs^2)^{-1/2}\,, \qquad F_0(s) = 1 - (1 - 4pqs^2)^{-1/2}\,.$$

Note that, for power series

$$P(s) = \sum_{n=0}^{\infty} a_n s^n$$

which converges for $|s| < 1$, we have that P being continuous on $(-1, 1)$. Then if $\sum_{n=0}^{\infty} a_n$ converges, then $\sum_{n=0}^{\infty} a_n = \lim_{s \to 1^-} P(s)$. Notice that $1 - 4pqs^2$, at $s = 1$, k attains its maximum at 1 when $p = 1/2$, and is non-negative. If T is the random variable indicating the time of the first return, then we have seen that $F_0(s) = G_T(s)$, and that we have

$$G_T(s) = \sum_{n=0}^{\infty} f_0(x)\, s^n = 1 - \left(1 - 4pqs^2\right)^{1/2} .$$

Furthermore, we see that $\sum_{n=0}^{\infty} f_0(n) \leq 1$, and we can write

$$\sum_{n=0}^{\infty} f_0(n) = \lim_{s \to 1^-} 1 - \left(1 - 4pqs^2\right)^{1/2} .$$

For the case where $p \neq q$, we have that

$$\sum_{n=0}^{\infty} f_0(n) = 1 - (1 - 4pq)^{1/2} < 1 \,,$$

thus the probability of the walker never return back to zero, in the case where $p \neq q$, or that $p \neq 1/2$, is given by

$$(1 - 4pq)^{1/2} = ((p+q)^2 - 4pq)^{1/2} = (p^2 + 2pq + q^2 - 4pq)^{1/2} = ((p-q)^2)^{1/2} = |p - q| \,.$$

For the case where $p = 1/2$, we have $q = 1/2$, and thus

$$|p - q| = 0 \,,$$

that is the probability return to zero at some time n in the future is exactly 1. Next we would like to find the expected time of return. Here ewe have

$$F_0'(s) = \sum_{n=0}^{\infty} f_0(n) n s^{n-1} \,.$$

Thus we have

$$F'(1) = \sum_{n=0}^{\infty} n f_0(n) = \mathbb{E}(T) \,.$$

In the case where $p = 1/2$, we have that

$$F_0(s) = 1 - (1 - s^2)^{1/2} \,,$$

thus we have

$$F_0'(s) = \frac{1}{2} \frac{2s}{(1 - s^2)^{1/2}} \,,$$

which implies

$$F_0'(1) = \lim_{s \to 1^-} \frac{2s}{(1 - s^2)^{1/2}} = \infty \,.$$

We have seen that, if a random walk is biased, say $q > p$ as probability going to the left. Then there is a positive probability that the walker will never visit the right hand side. Consider a random variable T_1, with takes value in $\mathbb{N}$, representing the first arrival at

$x = 1$, starting at $x = 0$. Let $F_1(s)$ denote the probability generating function of T_1, which reads

$$F_1(s) = \sum_{i\in\mathbb{N}} \mathbb{P}(T_1 = n)\, s^n = 0 + ps + \text{higher order terms}\,,$$

then we see here

$$F_1(1) = \sum_{n=1}^{\infty} \mathbb{P}(T_1 = n) \leq 1\,,$$

and from the previous result, we could have that $T_1 = \infty$ with nonzero probability, that is, we would like to show that $F_1(1) < 1$. In general, one is interested in finding

$$F_r(s) = \sum_{n\in\mathbb{N}} \mathbb{P}(T_r = n)\, s^n$$

where we have probability density function $f_r(n) = \mathbb{P}(\text{waker arrives at } r \text{ for the first time at time } n)$ for T_r. Note that we can write

$$F_r(s) = (F_1(s))^r\,, \tag{*}$$

which holds because, as one wants to arrive at r, one first needs to arrive at $r - 1$. For $n \geq 1$ and $r \geq 1$, we have

$$f_r(n) = \sum_{k=0}^{n} f_{r-1}(n-k)\, f_1(k)\,,$$

and multiplying both sides by s^n we obtain

$$f_r(n)\, s^n = \sum_{k=0}^{n} s^{n-k}\, f_{r-1}(n-k)\, s^k\, f_1(k)\,,$$

an inductive argument that yields (*) follows from here. Furthermore, we have that $f_1(0) = 0$, $f_1(1) = p$, $f_1(2) = 0$, and that

$$\begin{aligned} f_1(3) &= q\,\mathbb{P}(\{\text{first arrival at 1, starting from } -1\} = 2) \\ &= q\,\mathbb{P}(\{\text{first arrival at 2, starting from 0}\} = 2) = q\, f_2(2)\,. \end{aligned}$$

Thus in general, one can show that, for $n \geq 2$, we have

$$f_1(n) = q\, f_2(n-1)\,.$$

Then we can write

$$s^n\, f_1(n) = sq\, f_2(n-1)\, s^{n-1}\,,$$

and thus

$$F_1(s) = sp + sq\, F_2(s)\,.$$

Rearranging using (*), we obtain

$$sq\, F_1^2 - F_1 + sp = 0\,,$$

and solving we obtain

$$F_1(s) = \frac{1 - \sqrt{1 - 4pqs^2}}{2qs}\,.$$

Evaluating at $s = 1$, we see that

$$F_1(1) = \frac{1 - \sqrt{1 - 4pq}}{2q} = \frac{1 - \sqrt{(p+q)^2 - 4pq}}{2q} = \frac{1 - \sqrt{(p-q)^2)}}{2q} = \frac{1 - |p-q|}{2q}\,.$$

Thus we have

$$F_1(1) = \begin{cases} 1 & p > q \\ p/q & p < q \end{cases}\,.$$

Combining all, we conclude that we have

$$\mathbb{P}(T_1 = \infty) = 1 - \frac{p}{q} < 0\,, \qquad \text{when } p < q$$

and that is, the probability of the walker never visit the RHS is greater than zero.

Characteristic Functions

In this chapter, we will discuss the Law of Large Numbers, and the Central Limit Theorem. In order to get these, we have to understand the convergence of random variables, which involves the moment generating functions and their complex analogue, and the characteristic functions of random variables. We would like to translate properties of random variables to moment generating functions $M_X(t)$, or characteristic functions $\phi_X(t)$ which we will define next.

Note here, one can write

$$M_X(t) = \sum_n \mathbb{E}(X^n)\,\frac{t^n}{n!}\,,$$

where $\mathbb{E}(X^n)$ is called the n-th moment of X, but here $M_X(t)$ does not necessarily converge. Furthermore, there are random variables X, Y, with the properties that, all of their moments exist and $\mathbb{E}(X^n) = \mathbb{E}(Y^n)$, and thus $M_X(t) = M_Y(t)$, but at the same time their density function, or mass function, $f_X(x) \neq f_Y(x)$.

Definition 3.0.1
Here we define, for a continuous random variable X,

$$\phi_X(t) = \mathbb{E}(e^{itX}) = \int_{-\infty}^{\infty} e^{itx}\, f_X(x)\, dx\,.$$

$$M_X(t) = \mathbb{E}(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} f_X(x)\, dx\,.$$

In general, for a continuous random variable X, one finds that we have

$$\left| \int_{-\infty}^{\infty} e^{itx} \, f_X(x) \, dx \right| \leq \int_{-\infty}^{\infty} \left| e^{itx} \, f_X(x) \right| \, dx = \int_{-\infty}^{\infty} f_X(x) \, dx = 1 \, .$$

Thus we see here, $\phi_X(t)$, $|\phi_X(t)| \leq 1$, and $\phi_X(t)$ is uniformly continuous. That is, given t, if we have $\phi_{X_n}(t) \to \phi_X(t)$, then we have that $F_{X_n}(t) \to F_X(t)$, in which case we say X_n converges to X in distribution, denoted as $X_n \to_D X$. Furthermore, if F_X is continuous, then $F_{X_n} \to F_X$, that is $\mathbb{P}(X_n \leq c) \to \mathbb{P}(X \leq c)$.

Here we consider the Fourier inversion,

$$f_X(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-ixt} \, \phi_X(t) \, dt \, , \tag{*}$$

which holds when we have

$$\int_{-\infty}^{\infty} |\phi_X(t)| \, dt < \infty \, ,$$

and that $f_X(x)$ needs to be differentiable and $\phi_X(t)$ needs to be continuous at $t = 0$. These requirements to write (*) is assumed to be satisfied in general cases.

Example: Note that if one has $\phi_X(t) = 1$, then

$$1 = \int_{-\infty}^{\infty} e^{ixt} \, \delta_{x=0} \, dt = e^{i \cdot 0t} = 1 \, ,$$

where the Dirac-delta *function* $\delta_{x=0}$ is the density function of a singleton.

Example: Here we consider X_i to be i.i.d. $\sim X$, with $S_n = \sum_{i=1}^{n} X_i$. If X has sufficient moment, that is $\mathbb{E}(|X|^n) < \infty$, then by independence, we can write

$$\phi_{S_n}(t) = \phi_{X_1}(t) \cdot \phi_{X_2}(t) \cdot \cdots \cdot \phi_{X_n}(t) = (\phi_X(t))^n \, .$$

If one would like to scale the random variable X, that is we consider aX for some $a > 0$, then we can write

$$\phi_{aX}(t) = \int_{-\infty}^{\infty} e^{iaxt} f_X(x) \, dx = \phi_X(at) \, .$$

Now we can write

$$\phi_{S_n/n}(t) = \phi_{S_n}(t/n) = (\phi_X(t/n))^n \, .$$

Note that expanding $\phi_X(t)$ one would find

$$\phi_X(t) = 1 + i\mu t + (\text{higher order terms}) \, ,$$

where μ is the expectation of X. Thus we see here

$$\phi_{S_n/n}(t) = \left(1 + i\mu \frac{t}{n} + \underbrace{(\text{higher order terms in } t/n)}_{o(t/n)} \right)^n$$

thus taking the limit, and using the fact that

$$\lim_{n\to\infty}\left(1+\frac{a}{n}\right)^n = e^a\,,$$

we obtain

$$\lim_{n\to\infty}\phi_{S_n/n} = e^{i\mu t} = \phi_{Y=\mu}(t)\,,$$

and that for random variable $Y=\mu$, we have

$$\phi_Y(t) = M_Y(it) = \sum_n \frac{(i\mu t)^n}{n!} = e^{i\mu t}\,.$$

Thus by convergence theorem, we have that $F_{S_n/n} \to F_Y$, that is $S_n/n \to_D \mu$, a form of stating the law of large numbers. Notice on the other hand, naturally we can write

$$\mathbb{E}(S_n/n) = \mu\,, \qquad \mathrm{Var}(S_n/n) = \frac{1}{n^2}\mathrm{Var}(S_n) = \frac{1}{n}\mathrm{Var}(X) \to 0 \text{ when } n\to\infty\,.$$

Furthermore, one can define $Y_i = (X_i-\mu)/\sigma_{X_i}$, then we have

$$\mathbb{E}(Y_i) = \frac{\mathbb{E}(X_i-\mu)}{\sigma_{X_i}} = 0\,, \qquad \mathrm{Var}(Y_i) = \frac{\mathrm{Var}(X_i-\mu)}{\sigma^2_{X_i}} = \frac{\sigma^2_{X_i}}{\sigma^2_{X_i}} = 1\,.$$

Theorem 3.1 (The Central Limit Theorem)
Given a random variable X with mean μ and $\mathbb{E}(X^2)<\infty$, let X_i be i.i.d. $\sim X$. We define $Y_i = (X_i-\mu)/\sigma_{X_i}$, and consider

$$\frac{S_n}{\sqrt{n}} := \frac{1}{\sqrt{n}}\sum_{i=1}^n Y_i\,.$$

Then we have $(S_n/\sqrt{n}) \to_D Z$ for large enough n, where $Z\sim$ Normal$[0,1]$.

Proof. From the result above, we see that

$$\mathbb{E}(S_n/\sqrt{n}) = 0\,, \qquad \mathrm{Var}(S_n/\sqrt{n}) = \frac{\mathrm{Var}(S_n)}{n} = \frac{n}{n}\mathrm{Var}(Y_i) = 1\,.$$

Here we can write, as $\mathbb{E}(S_n)=0$,

$$\phi_{S_n/\sqrt{n}}(t) = \left(1 - \frac{1}{2}\left(\frac{t}{\sqrt{n}}\right)^2 + \text{higher order terms}\right)^n\,,$$

and taking the limit we obtain

$$\lim_{n\to\infty}\phi_{S_n/\sqrt{n}}(t) = \exp\left(-\frac{t^2}{2}\right)\,.$$

On the other hand, for the unit normal Z, we have

$$\phi_Z(t) = e^{i\mu_Z t - \sigma_Z^2 t^2/2} = e^{-t^2/2}\,,$$

where we have $\sigma_Z = 1$ and $\mu_Z = 0$, thus concluding we have that

$$\lim_{n\to\infty}\phi_{S_n/\sqrt{n}}(t) = \phi_Z(t)\,.$$

Thus we have

$$\frac{1}{\sqrt{n}} S_n \to_D Z\,.$$

This completes the proof. Note that, expanding we can write

$$\frac{\sum_{i=1}^{n}(X_i - \mu)}{\sqrt{n}\sigma_i} \to_D Z\,.$$

□

Example: To check whether a coin is biased. One can flip the coin n times and check where we get approximately $n/2$ heads. Quantifying the process, we define X_i i.i.d. ∼Bernoulli[1/2], and S_n ∼Binomial[$n, 1/2$]. We we have that $\mu(X_i) = 1/2$, and $\mathrm{Var}(X_i) = 1/4$. Thus we have

$$\frac{S_n - n/2}{\sqrt{n}/2} \to_D Z\,.$$

That is, suppose we have $n = 1000$, and $S_n = 550$ heads, then one is interested in the probability that $S_n \geq 550$.

Estimation of Outliers

The basis of statistical tests is to obtain an estimation of the outliers or the *core*.

Lemma 4.0.1
Consider we have a continuous function $h : \mathbb{R} \to [0, \infty)$, then we can write

$$\mathbb{P}(h(X) \geq a) \leq \frac{\mathbb{E}(h(x))}{a}\,.$$

Proof. Let A be the event of $\{h(x) \geq a\}$, and let I_A be the indicator random variable for A. Then we have that

$$I_A = \begin{cases} 1 & h(x) \geq a \\ 0 & \text{otherwise} \end{cases}\,.$$

Here we observe that we have

$$h(x) \geq a I_A\,,$$

then we can write

$$\mathbb{E}(h(x)) \geq \mathbb{E}(a I_A) = a \cdot \mathbb{E}(I_A) = a \cdot \mathbb{P}(A)\,,$$

giving the desired result. □

Corollary 4.0.1 (Markov's Inequality)
Let X be a non-negative random variable with finite expectation. Then we have

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

Note: Note that the statement of Corollary 3.1.1 has no restriction on the value of a. That is, we can have $a < \mu = \mathbb{E}(X)$, or $a \gg \mu$.

Corollary 4.0.2 (Chebychev's Inequality)
Let X be a random variable that has finite expectation μ and variance σ^2, then we have

$$\mathbb{P}(|X-\mu| \geq a) = \mathbb{P}(|X-\mu|^2 \geq a^2) \leq \frac{\mathbb{E}((X-\mu)^2)}{a^2} = \frac{\sigma^2}{a^2}\,.$$

That is, from Corollary 3.1.2 one can say

$$\mathbb{P}(|X-\mu| \geq a) \leq \frac{1}{n^2}$$

if one measures a in units of σ, that is

$$\mathbb{P}(|X-\mu| \geq n\sigma) \leq \frac{\sigma^2}{n^2\sigma^2} = \frac{1}{n^2}\,.$$

Note: Notice that all we have discussed so far are estimation of upper bounds on outliers, but the more important ones are the lower bounds, as they are heavily used in statistical tests.

Example: To determine whether or not a coin is fair,