# MATH 302 - INTRODUCTION TO PROBABILITY

## 概率论导论

**Author**
Wenyou (Tobias) Tian
田文友
University of British Columbia
英属哥伦比亚大学
2024

# Contents

# 1 Section 1

## 1.1 Sample Space and Probabilities

**Definition 1.1.** A ***sample point*** is a possible outcome, denoted as $\omega$.
A ***sample space*** is the set of all sample points, denoted as $\Omega$.

An ***event*** is a subset of $\Omega$, with $F$ representing the set of all possible events, we then have

$$|F| = 2^{|\Omega|}$$

A probability measure is a function where:

$$\mathbb{P} : F \to [0, 1]$$

such that for an event $A \in F$, $\mathbb{P}(A)$ means the probability of event $A$ occuring.
It is trivial that $\mathbb{P}(\varnothing) = 0$ and $\mathbb{P}(\Omega) = 1$, where $\forall A \in F, \mathbb{P}(A) \in [0, 1]$.

**Theorem 1.1.** If events $A_1, A_2, \ldots$ are pairwise disjoint, that is, $\forall i \neq j, A_i \cap A_j = \varnothing$, we have

$$\mathbb{P}(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} \mathbb{P}(A_i)$$

We then define

**Definition 1.2.** The triple $(\Omega, F, \mathbb{P})$ is called a ***probability space***.

## 1.2 Random sampling

***Sampling*** is choosing an object at random from a given set.

**Theorem 1.2.** If all outcomes are equally likely, if $|\Omega| < \infty$, then

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

where $A$ is an event.

## 1.3 Infinitely many outcomes

**Definition 1.3.** Sample spaces that are finite or countably infinite are ***discrete***.

When $\Omega$ is discrete, we have $\mathbb{P}(A) = \sum_{\omega \in A} \mathbb{P}(\{\omega\})$.
Uncountably infinite sample spaces can be the set $[0, 1]$. Notice that in this case, we **do not** have $\mathbb{P}(A) = \sum_{\omega \in A} \mathbb{P}(\{\omega\})$.

## 1.4 Consequences of Rules of Probability

**Definition 1.4.** Complement of a set $A$:

$$A^C = \{\omega \in \Omega \mid \omega \notin A\}$$

Union of two sets $A$ and $B$:

$$A \cup B = \{\omega \in \Omega \mid \omega \in A \vee \omega \in B\}$$

Intersection of two sets $A$ and $B$:

$$A \cap B = \{\omega \in \Omega \mid \omega \in A \wedge \omega \in B\}$$

Difference of two sets $A$ and $B$:

$$A - B = \{\omega \in \Omega \mid \omega \in A \wedge \omega \notin B\}$$

We then have the following properties:

1. $\mathbb{P}(A) = 1 - \mathbb{P}(A^C)$
2. If $A = \bigcup_{i=1}^{n} A_i$, and $A_i$ are pairwise disjoint, $\mathbb{P}(A) = \sum_i A_i$.
3. If $B \subset A$, $\mathbb{P}(B) \leq \mathbb{P}(A)$.
4. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$
5. $\mathbb{P}(A \cup B \cup C) = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) - \mathbb{P}(A \cap B) - \mathbb{P}(B \cap C) - \mathbb{P}(C \cap A) + \mathbb{P}(A \cap B \cap C)$

## 1.5 Random Variables

**Definition 1.5.** A ***random variable*** (r.v.) is a function from $\Omega$ to $\mathbb{R}$. If we define a random variable $X$, then we know

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

**Definition 1.6.** The ***probability distribution*** of a random variable $X$ is the collection of probabilities $\mathbb{P}(X \in B)$ for $B \subset \mathbb{R}$, where $X \in B$ is defined as

$$\{\omega \in \Omega : X(\omega) \in B\}$$

A discrete random variable takes value on a discrete set, then the ***probability mass function*** (p.m.f) of a discrete random variable $X$ is defined to be the collection of probabilities

$$p(k) = \mathbb{P}(X = k)$$

for all $k$ values $X$ may take. This implies: $\mathbb{P}(X \in B) = \sum_{k \in B} p(k)$

# 2 Section 2 - Conditional Probability and Independence

## 2.1 Conditional Probability

We write $\mathbb{P}(A|B)$ to be "the probability that event $A$ occurs **given** event $B$ occurs".

**Definition 2.1.** The probability that $A$ occurs given $B$ occurs is defined to be:

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

where $\mathbb{P}(AB)$ represents that both $A$ and $B$ occurs.

**Theorem 2.1.** $\mathbb{P}(A_1A_2\dots A_n) = \mathbb{P}(A_1)\times\mathbb{P}(A_2|A_1)\times\cdots\times\mathbb{P}(A_n|A_1A_2\dots A_{n-1})$. We also have:

$$\mathbb{P}(A) = \mathbb{P}(A|B)\cdot\mathbb{P}(B) + \mathbb{P}(A|B^C)\cdot\mathbb{P}(B^C)$$

Furthermore, if $B_1,\dots,B_n$ is a partition of $\Omega$, then

$$\mathbb{P}(A) = \sum_{i=1}^{n}\mathbb{P}(A|B_i)\mathbb{P}(B_i)$$

## 2.2 Bayes' Formula

**Theorem 2.2.** ***Bayes' Formula*** is the following:

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\cdot\mathbb{P}(B)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A|B)\cdot\mathbb{P}(B)}{\mathbb{P}(A|B)\cdot\mathbb{P}(B)+\mathbb{P}(A|B^C)\cdot\mathbb{P}(B^C)}$$

This is often used to represent situations like "false positives", where the table is the following:

| Test ↓ Actual → | Positive | Negative |
|---|---|---|
| Positive | True Positive | False Positive |
| Negative | False Negative | True Negative |

## 2.3 Independence

**Definition 2.2.** $A$ and $B$ are independent if and only if $\mathbb{P}(A) = \mathbb{P}(A|B)$, equivalently

$$\mathbb{P}(AB) = \mathbb{P}(A)\times\mathbb{P}(B)$$

If $A$ and $B$ are independent, then $A$ and $B^C$ are also independent. Mutually exclusive (disjoint) is not equivalent to independence. The independence between two events is essentially **proportional overlap**. For independence with more than 2 events,

say $A_1, \ldots, A_n$, they are independent if and only if for any set of indices $1 \leq i_1 < i_2 < \cdots < i_k \leq n$, we have

$$\mathbb{P}(A_{i_1} A_{i_2} \ldots A_{i_k}) = \mathbb{P}(A_{i_1}) \times \mathbb{P}(A_{i_2}) \times \cdots \times \mathbb{P}(A_{i_k})$$

If we have $A$, $B$ and $C$, and we only have

1. $\mathbb{P}(AB) = \mathbb{P}(A) \times \mathbb{P}(B)$
2. $\mathbb{P}(BC) = \mathbb{P}(B) \times \mathbb{P}(C)$
3. $\mathbb{P}(CA) = \mathbb{P}(C) \times \mathbb{P}(A)$

they are only ***pairwise independent***.

**Definition 2.3.** Random variables $X_1, \ldots, X_n$ are independent if and only if, $\forall B_1, \ldots, B_n \subset \mathbb{R}$,

$$\mathbb{P}(X_1 \in B_1, X_2 \in B_2, \ldots, X_n \in B_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \in B_i)$$

This implies that discrete random variables $X_1, \ldots, X_n$ are independent if and only if, $\forall k_1, \ldots, k_n \in \mathbb{R}$, we have

$$\mathbb{P}(X_1 = k_1, X_2 = k_2, \ldots, X_n = k_n) = \prod_{i=1}^{n} \mathbb{P}(X_i = k_i)$$

## 2.4 Independent Trials

A ***trial*** is a run of an experiment, where we we consider an experiment with two outcomes: 1 to denote success with probability $p$, and 0 to denote failure with probability $1 - p$.

**Definition 2.4.** A ***Bernoulli random variable*** with parameter $p$ satisfies:

$$X = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1 - p \end{cases}$$

This random variable represents 1 independent trial.
We denote this random variable to be $X \sim \text{Bern}(p)$

If we have $n$ trials, and let random variable $X$ be the number of successes, we have

**Definition 2.5.** A ***binomial random variable*** with parameter $n$ and $p$ satisfies:

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

for $k = 0, 1, \ldots, n$.
We denote this random variable to be $X \sim \text{Bin}(n, p)$

If $X_1, \ldots, X_n \sim \text{Bern}(p)$, and they are all independent, then

$$X_1 + X_2 + \cdots + X_n \sim \text{Bin}(n, p)$$

If we have unbounded trials, and let random variable $X$ be the number of trials until the 1st success, we have

**Definition 2.6.** A ***geometric random variable*** with parameter $p$ satisfies:

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

for $k = 1, 2, \ldots$ We denote this random variable to be $X \sim \text{Geom}(p)$

## 2.5 Conditional Independence

**Definition 2.7.** Events $A$ and $B$ are said to be ***conditionally independent*** given event $D$ if

$$\mathbb{P}(AB|D) = \mathbb{P}(A|D) \times \mathbb{P}(B|D)$$

# 3 Section 3 - Random Variables

## 3.1 Probability Distribution

Probability distribution for a random variable $X$ is the set of all probabilities $\{\mathbb{P}(X \in B)\}$, we use a p.m.f for discrete random variables to represent its probability distribution. What if the random variable is not discrete?

**Definition 3.1.** A random variable $X$ has ***probability density function*** (p.d.f) $f$ if

$$\mathbb{P}(X \leq a) = \int_{-\infty}^{a} f(x)\mathrm{d}x$$

If $X$ has a p.d.f, we call $X$ to be a ***continuous random variable***. A valid p.d.f must satisfy two conditions:

1. $\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$
2. $f(x) \geq 0$

We have the following properties of a p.d.f:

1. $\mathbb{P}(X \in [a, b]) = \int_a^b f(x)\mathrm{d}x$
2. $\mathbb{P}(X \in B) = \int_B f(x)\mathrm{d}x$
3. $\mathbb{P}(X = k) = 0$

**Definition 3.2.** Let random variable $X$ have p.d.f:

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

We say $X$ is a uniform variable with parameters $a, b$, where $X \sim \text{Unif}(a, b)$

The intuitive meaning of $f(x)$ is the following:

$$f(a) \approx \frac{\mathbb{P}(X \in [a, a + \varepsilon])}{\varepsilon}$$

## 3.2 Cumulative Distribution Function

**Definition 3.3.** The ***cumulative distribution function*** (c.d.f), $F$, of a random variable $X$ satisfies:

$$F(s) = \mathbb{P}(X \leq s)$$

for all $s \in \mathbb{R}$.
The c.d.f completely characterizes the probability distribution of a random variable.

Furthermore, if r.v. $X$ is continuous with p.d.f $f(x)$, we then have

$$F(s) = \int_{-\infty}^{s} f(x)\mathrm{d}x$$

This also gives that $f(x) = F'(x)$, for where $F'$ is defined, if $F'$ is undefined, then we choose an arbitrary value.

**Theorem 3.1.** If c.d.f is continuous and differentiable at all but finite number of points, then the underlying random variable is continuous.

The c.d.f of discrete r.v.s would have jumps at each available $X = k$.

**Theorem 3.2.** Suppose r.v. $X$ has c.d.f $F$ which is piecewise constant, then $X$ is a discrete r.v. The values that $X$ can take are the places where $F$ has jumps. If $x$ is such a point, then $\mathbb{P}(X = x)$ is the size of the jump.

We also know that a c.d.f must be **non-negative**, **always increasing**, and the limit as it approaches $\infty$ is 1.

## 3.3 Expectation

**Definition 3.4.** The expected value of a r.v. $X$ is:

1. Discrete: $\mathbb{E}(X) = \sum_k kp(k)$
2. Continuous: $\mathbb{E}(X) = \int_{-\infty}^{\infty} xf(x)\mathrm{d}x$

It is essentially the weighted average of values that $X$ can take.

Some expected values for common random variables are:

1. $X \sim \text{Bern}(p) : \mathbb{E}(X) = p$
2. $X \sim \text{Bin}(n, p) : \mathbb{E}(X) = np$
3. $X \sim \text{Unif}(a, b) : \mathbb{E}(X) = \frac{a+b}{2}$
4. $X \sim \text{Geom}(p) : \mathbb{E}(X) = \frac{1}{p}$

**Theorem 3.3.**

$$\mathbb{E}(X_1 + \cdots + X_n) = \mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n)$$

Specifically for a geometric r.v., we have the property:

$$\mathbb{E}(X) = \sum_{k\geq 1} \mathbb{P}(X \geq k) = \sum_{k\geq 1} (1-p)^{k-1}$$

This is derived from expected values of non-negative r.v.s where:

1. Discrete: $\mathbb{E}(X) = \sum_{k=1}^{\infty} \mathbb{P}(X \geq k)$
2. Continuous: $\mathbb{E}(X) = \int_0^{\infty} \mathbb{P}(X \geq x)\mathrm{d}x$

**Theorem 3.4.** Let the range of r.v. $X$ be contained in the domain of some real function $g$, then

$$\mathbb{E}(g(X)) = \sum_k g(k)p(k)$$

$$\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)\mathrm{d}x$$

**Definition 3.5.** The $n$th ***moment*** of r.v $X$ is defined to be $\mathbb{E}(X^n)$

## 3.4 Variance

**Definition 3.6.** Let $X$ be a r.v with mean $\mu$, then the ***variance*** of $X$ is

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

We may also denote it as $\sigma^2(X)$, where $\sigma(X)$ denotes the ***standard deviation*** of $X$ with $\sigma(X) = \sqrt{\mathrm{Var}(X)}$.

We can consider a function $g$, where $g(X) = (X - \mu)^2$. Then we may use Theorem 3.4 to yield corresponding formulas for discrete and continuous r.v.s.
Some variances of common r.v.s include:

1. $X \sim \mathrm{Bern}(p) : \sigma^2(X) = p(1-p)$
2. $X \sim \mathrm{Bin}(n, p) : \sigma^2(X) = np(1-p)$
3. $X \sim \mathrm{Geom}(p) : \mathbb{E}(X) = \frac{1-p}{p^2}$

**Theorem 3.5.** If $X$ is constant, then $\sigma^2(X) = 0$.
If $X_1, \ldots, X_n$ are independent, then $\sigma^2(X_1 + \cdots + X_n) = \sigma^2(X_1) + \cdots + \sigma^2(X_n)$.

$$\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$$

$$\sigma(aX + b) = a^2\sigma^2(X)$$

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

### 3.5 Gaussian Distribution

**Definition 3.7.** R.v. $Z$ has ***standard normal distribution*** if it has p.d.f:

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

We denote $Z \sim N(0,1)$

The corresponding c.d.f is thus:

$$\Phi(s) = \mathbb{P}(Z \leq s) = \int_{-\infty}^{s} \phi(x) \mathrm{d}x$$

$Z$ is called standard normal because $\mathbb{E}(Z) = 0, \sigma^2(Z) = 1$. For the entire family of normal r.v.s, just consider $X = \sigma Z + \mu$, then we would have $X \sim N(\mu, \sigma^2)$, with $\mathbb{E}(X) = \mu$ and $\sigma^2(X) = \sigma^2$.
Then for such a normal r.v. $X$, we have

$$F(s) = \mathbb{P}(X \leq s) = \mathbb{P}(Z \leq \frac{s-\mu}{\sigma}) = \Phi(\frac{s-\mu}{\sigma})$$

$$f(s) = \frac{\mathrm{d}}{\mathrm{d}s} F(s) = \frac{1}{\sigma} \cdot \phi(\frac{s-\mu}{\sigma}) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(s-\mu)^2}{2\sigma^2}}$$

If $X \sim N(\mu, \sigma^2)$ and $Y = aX + b$, then we have

$$Y \sim N(a\mu + b, a^2\sigma^2)$$

Furthermore, we have a special property for the standard normal c.d.f:

$$\Phi(t) = 1 - \Phi(-t)$$

# 4 Section 4

## 4.1 Law of Large Numbers (LLN), Binomial Version

Consider $X_i \sim \text{Bern}(p)$, and $S_n = X_1 + \cdots + X_n$ all of which are independent. We already know $S_n \sim \text{Bin}(n, p)$, but as $n$ gets large, it becomes very hard to calculate, thus we use binomial approximation:

1. First order approximation: $S_n \approx \mathbb{E}(S_n) = np$
2. Second order approximation (deviation from mean is Gaussian):

$$S_n \approx \mathbb{E}(S_n) + \sqrt{\sigma^2(S_n)} \cdot N(0,1) = np + \sqrt{np(1-p)} \cdot N(0,1)$$

**Theorem 4.1.** The ***Law of Large Numbers*** claim:

$$\forall \varepsilon > 0, \lim_{n\to\infty} \mathbb{P}(|\frac{S_n}{n} - p| < \varepsilon) = 1$$

This is equivalent to saying:

$$\forall \delta > 0, \exists N, n \geq N, 1 - \delta \leq \mathbb{P}(|\frac{S_n}{n} - p| < \varepsilon) \leq 1$$

We can call this $\frac{S_n}{n}$ converges to $p$ in probability.

We define $\frac{S_n}{n}$ to be ***proportion of success***, with expected value $p$.
Note that for any r.v. $X$, $\{|X| \leq a\} \subset \{X \leq a\}$.

## 4.2 Central Limit Theorem (CLT), Binomial Version

Let r.v. $S_n \sim \text{Bin}(n, p)$ with $\mathbb{E}(S_n) = np$ and $\sigma^2(S_n) = np(1-p)$, we standardize it to another r.v. $Q_n$ with:

$$Q_n = \frac{S_n - np}{\sqrt{np(1-p)}}$$

so $\mathbb{E}(Q_n) = 0$ and $\sigma^2(Q_n) = 1$.
Then $Q_n$ converges to $N(0,1)$ in distribution.

**Theorem 4.2.** The ***Central Limit Theorem*** (CLT)) claim:

$$\lim_{n\to\infty} \mathbb{P}(Q_n \in [a, b]) = \Phi(b) - \Phi(a)$$

The normal approximation goes from CLT saying:

$$\frac{S_n - np}{\sqrt{np(1-p)}} \to_d N(0,1)$$

This means for large $n$, $S_n$ is approximately $N(np, np(1-p))$, the approximation is accurate if $np(1-p) > 10$.

Consider a random walk problem, where one takes 1 step left or right each with $\frac{1}{2}$ probability, about how far is the person from home after $n$ steps, assuming the person starts from the house.
Let

$$X_i = \begin{cases} 1 & \frac{1}{2} \\ -1 & \frac{1}{2} \end{cases}$$

and $Y_i \sim \text{Bern}(\frac{1}{2})$, then $X_i = (Y_i - \frac{1}{2}) \times 2$. Then if we let $Z_n$ be the distance after $n$ steps, we have

$$Z_n = \sum_{i=1}^{n} X_i = 2\sum_{i=1}^{n}(Y_i - \frac{1}{2}) = -n + 2\sum_{i=1}^{n} Y_i = -n + 2S_n$$

Then by normal approximation, we have $Z_n \approx -n + 2 \cdot N(\frac{n}{2}, \frac{n}{4}) = -n + 2(\frac{n}{2} + \frac{\sqrt{n}}{2} \cdot N(0,1)) = \sqrt{n} \cdot N(0,1)$.
The concentration of standard normal r.v. is as follows:

1. $\mathbb{P}(N(0,1) \in [-1,1]) \approx 0.68$
2. $\mathbb{P}(N(0,1) \in [-2,2]) \approx 0.95$
3. $\mathbb{P}(N(0,1) \in [-3,3]) \approx 0.997$

We then know with normal approximation:

$$\text{Bin}(n,p) \in [np - 3\sqrt{np(1-p)}, np + 3\sqrt{np(1-p)}]$$

## 4.3 Application of Normal Approximation and Confidence Intervals

To find a 95% confidence interval, we first start with an estimator $\hat{\mu}$. We express this estimator as some normal r.v. with information about $\mu$. We then find the distribution of $\hat{\mu} - \mu$ and build relationship with a standard normal distribution $N(0,1)$. Finally, we may use the concentration of a standard normal variable to find the 95% confidence interval.
Recall with normal approximation $\mathbb{P}(\text{Bin}(n,p) \in [a,b]) \approx \mathbb{P}(N(np, np(1-p)) \in [a,b])$, then for large $n$, we should have for $Z \sim N(0,1)$,

$$S_n \approx np + \sqrt{np(1-p)} \cdot Z$$

and $\hat{p} = \frac{S_n}{n}$. So $\hat{p} - p \approx \sqrt{\frac{p(1-p)}{n}} \cdot Z$.

$$\begin{aligned} \mathbb{P}(|\hat{p} - p| \le \varepsilon) &\approx \mathbb{P}(|Z| \le \frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}) \\ &\ge \mathbb{P}(|Z| \le 2\varepsilon\sqrt{n})\{\sqrt{p(1-p)} \le \frac{1}{2}\} \end{aligned}$$

For a 95% confidence interval, we choose $\varepsilon = \frac{1}{\sqrt{n}}$, then we know:

$$p \in [\hat{p} - \frac{1}{\sqrt{n}}, \hat{p} + \frac{1}{\sqrt{n}}]$$

## 4.4 Poisson Random Variable

**Definition 4.1.** A discrete r.v. $X$ has ***Poisson*** distribution with parameter $\lambda > 0$ if:

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

We denote $X \sim \text{Poisson}(\lambda)$.

**Theorem 4.3.** Let $\lambda > 0$, $k \in \mathbb{Z}^+$, then:

$$\lim_{n \to \infty} \mathbb{P}(\text{Bin}(n, \frac{\lambda}{n}) = k) = \mathbb{P}(\text{Poisson}(\lambda) = k)$$

The p.m.f of $\text{Bin}(n, \frac{\lambda}{n})$ converges to the p.m.f of $\text{Poisson}(\lambda)$.
$\text{Bin}(n, p) \approx \text{Poisson}(np)$ for $n$ large and $p$ small.

**Theorem 4.4.** Let $X \sim \text{Bin}(n, p)$, and $Y \sim \text{Poisson}(np)$, then for any $A \subset \mathbb{Z}^+$, we have

$$|\mathbb{P}(X \in A) - \mathbb{P}(Y \in A)| \leq np^2$$

The properties of a Poisson r.v. include:

Let $X \sim \text{Poisson}(\lambda)$, $Y \sim \text{Poisson}(\alpha)$ 1. $\mathbb{E}(X) = \lambda, \sigma^2(X) = \lambda$
2. $X + Y \sim \text{Poisson}(\lambda + \alpha)$

Often it is natural to model as Poisson even without knowing $n$ or $p$ for underlying Binomial distribution. We only need to know $np = \lambda$.

## 4.5 Exponential Distribution

An exponential r.v. models continuous waiting time, analagous to geometric r.v. modelling discrete waiting time.

**Definition 4.2.** A continuous r.v. $X$ has ***exponential distribution*** with parameter $\lambda > 0$ if its p.d.f, $f$, is:

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

We write $X \sim \text{Exp}(\lambda)$

Thus, we have $\mathbb{P}(X > t) = e^{-\lambda t}$, giving us the c.d.f, $F$, to be:

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

A useful property is that $X \sim \text{Exp}(\alpha), Y \sim \text{Exp}(\beta)$, and they are independent, if we define $Z = \min(X, Y)$, then $Z \sim \text{Exp}(\alpha + \beta)$.
Other general properties include:

1. $\mathbb{E}(X) = \frac{1}{\lambda}, \sigma^2(X) = \frac{1}{\lambda^2}$
2. $\alpha X \sim \text{Exp}(\frac{\lambda}{\alpha})$
3. Memorylessness: $\mathbb{P}(X > s + t \mid X > t) = \mathbb{P}(X > s)$

# 5 Section 5

## 5.1 Moment Generating Function

**Definition 5.1.** The ***moment generating function*** of a r.v. $X$ is the function $M : \mathbb{R} \to \mathbb{R}^+$ defined by:

$$M(t) = M_X(t) = \mathbb{E}(e^{tX})$$

The m.g.f of some common r.v.s include:

1. $X \sim \text{Bern}(p)$: $M(t) = 1 - p + pe^t$
2. $X \sim \text{Unif}(0, 1)$: $M(t) = \frac{1}{t}(e^t - 1)$
3. $X \sim N(0, 1)$: $M(t) = e^{\frac{1}{2}t^2}$
4. $X \sim \text{Exp}(\lambda)$:

$$M(t) = \begin{cases} \frac{\lambda}{\lambda - t} & t < \lambda \\ \infty & t \geq \lambda \end{cases}$$

**Theorem 5.1.** M.g.f can generate moments in the following way:

$$M^{(n)}(0) = \mathbb{E}(X^n)$$

That is

$$\frac{\mathrm{d}^n}{\mathrm{d}t^n} M(t)|_{t=0} = \mathbb{E}(X^n)$$

We say that $X$ and $Y$ are equal in distribution if $\forall B \subset \mathbb{R}$, $\mathbb{P}(X \in B) = \mathbb{P}(Y \in B)$.

**Theorem 5.2.** If $M_X(t) = M_Y(t) \neq \infty$, we say $X \overset{\text{dist.}}{=} Y$.
More generally, if $\exists \delta > 0$, $\forall t \in (-\delta, \delta)$, $M_X(t) = M_Y(t) \neq \infty$, we say $X \overset{\text{dist.}}{=} Y$.

## 5.2 Distribution of a Function of a Random Variable

**Theorem 5.3.** Let $X$ be a discrete r.v. with p.m.f $p_X$, $g$ be an one-to-one function with $Y = g(X)$, then $Y$ has p.m.f satisfying $p_Y(g(k)) = p_X(k)$

If $X$ is still discrete, but $g$ is no longer $one - to - one$, then we have

$$p_Y(l) = \sum_{k:g(k)=l} p_X(k)$$

For continuous random variables, we derive its c.d.f for $g(X)$.
For affine transformations, that is $Y = aX + b$, we should have

$$f_Y(s) = \frac{1}{|a|} \times f_X(\frac{s - b}{a})$$

If $g$ is one-to-one and differentiable for continuous r.v. $X$, and $Y = g(X)$, we would have:

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

We also need $g$ have derivative at only finite number of points and $g^{-1}$ existing. For other points, $f_Y(y) = 0$. If $g$ is only differentiable and has derivative at finitely many number of points, with $Y = g(X)$, we have:

$$f_Y(y) = \sum_{x:g(x)=y, g'(x)\neq 0} \frac{f_X(x)}{|g'(x)|}$$

# 6 Section 6 - Joint Distribution of Random Variables

## 6.1 Joint Distribution of Discrete Random Variables

**Definition 6.1.** Let $X, Y$ be discrete r.v.s defined on the same sample space, then the ***joint p.m.f*** is defined by:

$$p(k,l) = \mathbb{P}(X = k, Y = l)$$

for all values of $k$ and $l$.

**Definition 6.2.** Given a joint p.m.f of $p(k,l)$, the p.m.f of $X$ is defined by:

$$p_X(k) = \sum_l p(k,l)$$

where $p_X$ is the ***marginal p.m.f*** of $X$

**Theorem 6.1.** Let $g : \mathbb{R}^2 \to \mathbb{R}$, and let $X, Y$ be discrete r.v. with joint p.m.f $p$, then:

$$\mathbb{E}(g(X,Y)) = \sum_{k,l} g(k,l)p(k,l)$$

A famous example in game theory is the situation of Prisoner's Dilemma.

## 6.2 Jointly Continuous Random Variables

**Definition 6.3.** R.v.s $X, Y$ are ***jointly continuous*** if $\exists f : \mathbb{R}^2 \to \mathbb{R}^+$ such that $\forall B \subset \mathbb{R}^2$, we have

$$\mathbb{P}((X,Y) \in B) = \iint_B f(x,y)\mathrm{d}x\mathrm{d}y$$

We need:

1. $f(x,y) \geq 0$
2. $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\mathrm{d}x\mathrm{d}y = 1$

**Definition 6.4.** Let $g : \mathbb{R}^2 \to \mathbb{R}$, the expected value of $g$ can be characterized by:

$$\mathbb{E}(g(X,Y)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f(x,y)\mathrm{d}x\mathrm{d}y$$

**If $(X,Y)$ are jointly continuous, then $X, Y$ are continuous r.v.s**. However, if $X, Y$ are continuous, $(X,Y)$ need not be jointly continuous.

**Definition 6.5.** Let $(X, Y)$ be jointly continuous with joint p.d.f $f$, then $X$ is continuous with p.d.f $f_X$ satisfying:

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\mathrm{d}y$$

We specifically analyze uniform r.v.s in higher dimensions. Consider $D \subset \mathbb{R}^2$ with $A_D < \infty$, then $(X, Y)$ has uniform distribution on $D$ if:

$$f(x, y) = \begin{cases} \frac{1}{A_D} & (x, y) \in D \\ 0 & (x, y) \notin D \end{cases}$$

Let $(X_1, \ldots, X_n) \sim \mathrm{Unif}(\sqrt{n}B_n)$, with $B_n = \{(x_1, \ldots, x_n) : x_1^2 + \cdots + x_n^2 \leq n\}$, then we first have:

$$f(x_1, \ldots, x_n) = \begin{cases} \frac{1}{\mathrm{Vol}(\sqrt{n}B_n)} & (x_1, \ldots, x_n) \in \sqrt{n}B_n \\ 0 & (x_1, \ldots, x_n) \notin \sqrt{n}B_n \end{cases}$$

For $r > 0$, we also have $\mathrm{Vol}(rB_n) = r^n \mathrm{Vol}(B_n)$.
The marginal of $X_1$ is derived to be ($V_i = \mathrm{Vol}(B_i)$):

$$f_{X_1}(x_1) = \frac{V_{n-1} \cdot n^{\frac{n}{2}}}{V_n \sqrt{n}^n \sqrt{n - x_1^2}} \cdot (\frac{n - x_1^2}{n})^{\frac{n}{2}}$$

when taking $n \to \infty$, it converges to $\phi(x_1)$

## 6.3 Joint Distribution and Independence

For $X, Y$ be discrete r.v.s with joint p.m.f $p$ and marginal $p_X$, $p_Y$, $X$ and $Y$ are independent if and only if:

$$p(k, l) = p_X(k) \times p_Y(l)$$

Analagously, for $X, Y$ be continuous r.v.s with joint p.d.f $f$ and marginal $f_X$, $f_Y$, $X$ and $Y$ are independent if and only if:

$$f(x, y) = f_X(x) \times f_Y(y)$$

# 7 Section 7

## 7.1 Sums of Independent Random Variables

**Definition 7.1.** Let $X, Y$ be independent r.v.s, we have $X+Y$ satisfying:

1. Discrete: $p_{X+Y}(n) = \sum_k p_X(k)p_Y(n-k) = p_X * p_Y(n) = p_Y * p_X(n)$
2. Continuous: $f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(x)f_Y(z-x)\mathrm{d}x = f_X * f_Y(z)$

Here $p_X * p_Y$ is the ***convolution*** of $p_X$ and $p_Y$, $f_X * f_Y$ is the ***convolution*** of $f_X$ and $f_Y$.

# 8 Section 8 - Expectation and Variance in Multivariate Setting

## 8.1

## 8.2 Expectation and Variance for Sums of r.v.s

**Theorem 8.1.** Regardless of $X, Y$ being independent or not, we have:

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

$$\mathbb{E}(f(X) + g(Y)) = \mathbb{E}(f(X)) + \mathbb{E}(g(Y))$$

For $X, Y$ being independent r.v.s, we have:

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

$$\sigma^2(X+Y) = \sigma^2(X) + \sigma^2(Y)$$

$$\mathbb{E}(\frac{X}{Y}) = \mathbb{E}(X) \cdot \mathbb{E}(\frac{1}{Y})$$

**Theorem 8.2.** Let $X_1, \ldots, X_n$ be independently identically distributed r.v.s with $\mathbb{E}(X_i) = \mu$ and $\sigma^2(X_i) = \sigma^2$, if we define:

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n}$$

we have $\mathbb{E}(\bar{X}_n) = \mu$ and $\sigma^2(\bar{X}_n) = \frac{\sigma^2}{n}$

**Definition 8.1.** Let $A$ be an event, the indicator of that event $1_A$ is a random variable satisfying:

$$1_A = \begin{cases} 1 & \text{A happens} \\ 0 & \text{A does not happen} \end{cases}$$

We know that $1_A \sim \text{Bern}(p)$, with $p = \mathbb{P}(A)$ and $\mathbb{E}(1_A) = \mathbb{P}(A)$

## 8.3 Sums and Moment Generating Functions

We know functions of $X, Y$ are independent if $X, Y$ are independent. Then, if $X, Y$ are independent, $M_{X+Y}(t) = M_X(t)M_Y(t)$. That is the m.g.f of sum is product of m.g.f.
For $X \sim N(0, \sigma^2)$, we have $M_X(t) = e^{\frac{t^2\sigma^2}{2}}$.
Since $X \sim \text{Bin}(n, p)$ and we know $X \overset{\text{dist.}}{=} X_1 + \cdots + X_n$ with $X_i \sim \text{Bern}(p)$, so we know

$$M_X(t) = (1 - p + pe^t)^n$$

For Poisson variables $X \sim \text{Poisson}(\alpha)$, we have

$$M_X(t) = \exp\big(\alpha(e^t - 1)\big)$$

And in this case, if $X_1, \ldots, X_n \sim \text{Poisson}(1)$, we know $\sum_i X_i \sim \text{Poisson}(n)$

## 8.4 Covariance and Correlation

To say something about the level of dependence between $X$ and $Y$, we can look into the covariance of $X$ and $Y$.

**Definition 8.2.** Let $X, Y$ be r.v.s with $\mu_X, \mu_Y$, we have

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

If $X$ and $Y$ are independent, then the covariance of $X$ and $Y$ is 0. However, the other way may not stand, since if $f$ is an even function, $X \sim (-a, a)$, then $\text{Cov}(X, f(X)) = 0$.
The properties of covariances include:

1. $|\text{Cov}(X, Y)| \leq \sqrt{\sigma^2(X)\sigma^2(Y)}$
2. $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
3. $\text{Cov}(aX + bZ, Y) = a\text{Cov}(X, Y) + b\text{Cov}(Z, Y)$
4. $\text{Cov}(X, X) = \sigma^2(X)$

Consider two events $A, B$ and their corresponding indicator variables $1_A, 1_B$, then we have

$$\text{Cov}(1_A, 1_B) = \mathbb{E}(1_A 1_B) - \mathbb{E}(1_A)\mathbb{E}(1_B) = \mathbb{P}(AB) - \mathbb{P}(A)\mathbb{P}(B)$$

Thus, $\text{Cov}(1_A, 1_B) = 0 \iff A, B$ are independent.

**Theorem 8.3.** For r.v.s $X_1, \ldots, X_n$, we have

$$\sigma^2(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \sigma^2(X_i) + 2 \sum_{1 \leq i < j \leq n} \text{Cov}(X_i, X_j)$$

**Definition 8.3.**

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\sigma^2(X)\sigma^2(Y)}}$$

We also know it is $\in [-1, 1]$

Then $Y = aX + b \iff \text{Corr}(X, Y) = 1$.
Furthermore, if $X_i$ and $X_j$ are pairwise uncorrelated, then we still have:

$$\sigma^2(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \sigma^2(X_i)$$

# 9 Section 9

## 9.1 Estimating Tail Probabilities

**Theorem 9.1.** ***Markov Inequality*** states that, for a non-negative random variable $X$ and some $t > 0$, we have

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}(X)}{t}$$

The proof relies on $X \geq t \cdot 1_{[X \geq t]}$.

**Theorem 9.2.** ***Chebychev's Inequality*** states that, for a random variable $X$ with $\mathbb{E}(X) = \mu$ and $\sigma^2(X) = \sigma^2$, then for some $t > 0$, we have:

$$\mathbb{P}(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}$$

For $X_i$ be uncorrelated (independence suffices) with $\mathbb{E}(X_i) = \mu$ and $\sigma^2(X_i) = \sigma^2$, then define $S_n = X_1 + \cdots + X_n$, we have

$$\mathbb{P}(|S_n - n\mu| \geq \sqrt{n}t) \leq \frac{\sigma^2}{t^2}$$

$$\mathbb{P}(|\frac{S_n}{n} - \mu| \geq \frac{t}{\sqrt{n}}) \leq \frac{\sigma^2}{t^2}$$

## 9.2

## 9.3 Law of Large Numbers, Central Limit Theorem, General Version

**Theorem 9.3.** Let $X_1, \ldots, X_n$ be iid with $\mathbb{E}(X_i) = \mu, \sigma^2(X_i) = \sigma^2 < \infty$, let

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n}$$

then $\forall \varepsilon > 0$,

$$\lim_{n \to \infty} \mathbb{P}(|\bar{X}_n - \mu| \leq \varepsilon) = 1$$

$$\lim_{n \to \infty} \mathbb{P}(|\bar{X}_n - \mu| > \varepsilon) = 0$$

If we only focus on a small interval around $\mu$, we ask about the distribution of $\bar{X}_n$. We examine $Z_n = \frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu)$, so that $\mathbb{E}(Z_n) = 1$ and $\sigma^2(Z_n) = 1$.

**Theorem 9.4.** Let $X_1, \ldots, X_n$ be iid with $\mathbb{E}(X_i) = \mu, \sigma^2(X_i) = \sigma^2 < \infty$, let

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

Then $Z_n \to_d N(0,1)$, $\forall a \leq b \in (-\infty, \infty)$, we have

$$\lim_{n\to\infty} \mathbb{P}(Z_n \in [a,b]) = \Phi(b) - \Phi(a)$$

This means a normalized sum of many iid r.v.s is almost $N(0,1)$, that is

$$\frac{1}{\sqrt{n}} \sum \pm 1 \approx N(0,1)$$

The normal approximation start with $X_1, \ldots, X_n$ being iid, with $\mathbb{E}(X_i) = \mu$ and $\sigma^2(X_i) = \sigma^2$, let $S_n = X_1 + \cdots + X_n$, thus, we know $\mathbb{E}(S_n) = n\mu$ and $\sigma^2(S_n) = n\sigma^2$, so if $n$ is large

$$S_n \approx N(n\mu, n\sigma^2) = n\mu + \sqrt{n}\sigma N(0,1)$$

**Theorem 9.5.**

$$|\mathbb{P}(Z_n \leq x) - \Phi(x)| \leq \frac{3\mathbb{E}|X - \mu|^3}{\sigma^3\sqrt{n}}$$

The tail decay for normal random variables has:

$$\mathbb{P}(N(0,1) \geq t) \leq e^{-\frac{t^2}{2}}$$

# 10 Section 10

## 10.1 Conditional Distribution of Discrete Random Variables

**Definition 10.1.** Let $X$ be a discrete r.v., $B$ be an event with $\mathbb{P}(B) > 0$, the conditional p.m.f of $X$ given $B$ is defined as:

$$p_{X|B}(k) = \mathbb{P}(X = k \mid B) = \frac{\mathbb{P}(X = k, B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B \mid X = k)\mathbb{P}(X = k)}{\mathbb{P}(B)}$$

**Definition 10.2.** Following the same setup, the expected value of $X$ given $B$ is

$$\mathbb{E}(X \mid B) = \sum_k k \times p_{X|B}(k)$$

**Theorem 10.1.** The ***Law of Total Probability*** states that, if $B_1, \ldots, B_n$ is a partition of $\Omega$, then

$$\mathbb{P}(A) = \sum_i \mathbb{P}(A \mid B_i)\mathbb{P}(B_i)$$

This gives us

$$p_X(k) = \sum_{i=1}^{n} p_{X|B_i}(k) \cdot \mathbb{P}(B_i)$$

bringing the expected value to be

$$\mathbb{E}(X) = \sum_{i=1}^{n} \mathbb{E}(X \mid B_i) \cdot \mathbb{P}(B_i)$$

We then transition from conditioning on events to conditioning on r.v.s. Then,

**Definition 10.3.** Let $Y$ be another discrete r.v., so conditioning on $\{Y = y\}$, then

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

So this gives the conditional expected value of $X$ conditional on $Y = y$, we have

$$\mathbb{E}(X \mid Y = y) = \sum_x x \cdot p_{X|Y}(x|y)$$

Since $\{Y = y\}$ form a partition, then we have:

$$p_X(x) = \sum_y p_{X|Y}(x|y)p_Y(y)$$

$$\mathbb{E}(X) = \sum_y \mathbb{E}(X \mid Y = y) \cdot p_Y(y)$$

By Bayes' Formula, we also have:

$$p_{X|Y}(x|y) = \frac{p_{Y|X}(y|x)p_X(x)}{p_Y(y)}$$

## 10.2 Conditional Distribution for Continuous Random Variables

Analagous to discrete cases,

**Definition 10.4.** Let $X, Y$ be jointly continuous, the conditional function of $X$ given $Y = y$ is defined as:

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

The ***conditional expectation*** of $g(X)$ given $Y = y$ is defined as:

$$\mathbb{E}(g(X) \mid Y = y) = \int_{-\infty}^{\infty} g(x) \cdot f_{X|Y}(x|y)\mathrm{d}x$$

The ***conditional probability*** of $X \in A$ given $Y = y$ is defined as:

$$\mathbb{P}(X \in A \mid Y = y) = \int_A f_{X|Y}(x|y)\mathrm{d}x$$

Then by definition,

$$f_X(x) = \int_{-\infty}^{\infty} f_{X|Y}(x|y) \cdot f_Y(y)\mathrm{d}y$$

$$\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} \mathbb{E}(g(X) \mid Y = y) \cdot f_Y(y)\mathrm{d}y$$

## 10.3 Conditional Expectation

From previous definitions, we can consider $\mathbb{E}(X \mid Y = y) = v(y)$ to be a function of $y$. Then we transform this to $v(Y)$ so that $\mathbb{E}(X \mid Y) = v(Y)$ and it becomes a r.v.

**Theorem 10.2.** The ***Law of Iterated Expectation*** states that:

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

Conditional expectation respects linearity, thus, we have

$$\mathbb{E}(X_1 + X_2 \mid Y) = \mathbb{E}(X_1 \mid Y) + \mathbb{E}(X_2 \mid Y)$$

Independence means that conditioning has no effect, thus, we should have

$$p_{X|Y}(x|y) = p_X(x)$$

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

if and only if $X, Y$ are independent. This also adds that:

$$\mathbb{E}(X \mid Y = y) = \mathbb{E}(X), \mathbb{E}(X \mid Y) = \mathbb{E}(X)$$

Some properties include:

1. $\mathbb{E}(X \mid X = x) = x$
2. $\mathbb{E}(g(X) \mid X = x) = g(x)$
3. $\mathbb{E}(X \mid X) = X$
4. $\mathbb{E}(g(X) \mid X) = g(X)$
5. $\mathbb{E}(g(X)f(Z) \mid X) = g(X) \cdot \mathbb{E}(f(Z) \mid X)$