Queens' College Cambridge

Introduction to Probability

Alistair O'Brien

Department of Computer Science

July 7, 2021

Contents

1 Introduction

1.1 Combinatorics

1.1.1 Principle of Counting

Theorem 1.1.1. (**Addition Principle of Counting**) Suppose that we have two disjoint tasks T_1 and T_2 with n_1 ways of performing T_1 and n_2 ways of performing T_2.

The number of ways of performing T_1 **or** T_2 is $n_1 + n_2$.

Corollary 1.1.1.1. (**Generalised Addition Principle of Counting**) Suppose we have the disjoint tasks $T_1, T_2, \ldots, T_n$ with n_1 ways of performing T_1, and so on...

Then the number of ways of performing T_1 or T_2 or ... or T_n is $\sum_{i=1}^{n} n_i$.

Proof. We proceed by induction on n with the statement

$$P(n) = \forall \text{ disjoint tasks } T_1, \ldots, T_n.$$
$$\#\text{ of ways of performing } T_1 \text{ or } \cdots T_n = \sum_{i=1}^{n} n_i$$

and a basis of $n = 2$.
Base Case. $P(2)$ holds by Theorem ??.
Inductive Step. We wish to show that $\forall n \in \mathbb{N}_{\geq 2}.P(n) \implies P(n+1)$. Let $n \in \mathbb{N}_{\geq 2}$ be an arbitrary natural number. Let us assume that $P(n)$ holds, that is to say suppose $T_1, \ldots, T_n$ are some arbitrary disjoint tasks, then

$$\#\text{ of ways of performing } T_1 \text{ or } \cdots T_n = \sum_{i=1}^{n} n_i.$$

We wish to show that $P(n+1)$ holds. Let us consider the disjoint tasks $T_1, \ldots, T_{n+1}$. Let T denote the task of performing T_1 or ... or T_n. By our inductive hypothesis, it follows that the number of ways of performing T is $\sum_{i=1}^{n} n_i$. Instantiating theorem ?? it follows that the number of ways of performing T or T_{n+1} is

$$\text{\# of ways of performing } T \text{ or } T_{n+1} = \sum_{i=1}^{n} n_i + n_{n+1} = \sum_{i=1}^{n+1} n_i.$$

So we have $P(n+1)$.

By the Principle of Mathematical Induction, we conclude that $P(n)$ holds for all $n \in \mathbb{N}_{\geq 2}$. □

Theorem 1.1.2. (**Product Principle of Counting**) Suppose we have two tasks T_1 and T_2 with n_1 ways of performing T_1 and n_2 ways of performing T_2.

The number of ways of performing T_1, T_2 in sequence is $n_1 \cdot n_2$.

Corollary 1.1.2.1. (**Generalised Product Principle of Counting**) Suppose we have the tasks $T_1, \ldots, T_n$ are to be performed in sequence, with n_1 ways of performing T_1, and so on ... The number of ways of performing the sequence $T_1 T_2 \cdots T_n$ is $\prod_{i=1}^{n} n_i$.

Proof. Induction on n with a basis of $n = 2$. □

Theorem 1.1.3. (**The Pigeonhole Principle**) Suppose n pigeons are assigned to m pigeonholes and $m < n$, then at least one pigeonhole contains two or more pigeons.

Proof. We proceed by contradiction. Let us assume that each pigeonhole contains at most 1 pigeon. Since n pigeons are assigned to m pigeonholes and $m < n$, then not all the pigeons have been assigned. A contradiction! □

Theorem 1.1.4. (**The Extended Pigeonhole Principle**) Suppose there are m objects placed into n pigeonholes, then at least one pigeonhole has at least $\lceil \frac{m}{n} \rceil$ objects.

1.1.2 Permutations

Definition 1.1.1. (**Permutation**) A permutation of a set S is a ordered sequence $\pi = \langle x_1, \ldots x_n \rangle \subseteq S$

Theorem 1.1.5. Suppose $0 \leq k \leq n$, let ${}_nP_k$ denote the number of k element permutations of a set of n elements. Then

$${}_nP_k = n \cdot (n-1) \cdot (n-2) \cdots (n-k+1) = \frac{n!}{(n-k)!}.$$

Proof. Suppose we have a set S of n elements. Let $0 \leq k \leq n$ be an arbitrary integer. We have k tasks in sequence:

- T_1: Choose an element x_1 from S.
- T_2: Choose an element x_2 from $S \setminus \{x_1\}$.
- $\vdots$
- T_k: Choose an element x_k from $S \setminus \{x_1, \ldots, x_{k-1}\}$

Note that there are $|R|$ ways of choosing an element from a set T, so by the generalised product principle of counting, we have

$${}_nP_k = n \cdot (n-1) \cdots (n-k+1) = \frac{n!}{(n-k)!}.$$

□

Theorem 1.1.6. (**Permutations of Indistinct Objects**) The number of distinguishable permutations formed from a multi-set of n element where n_1 objects are indistinct from each other, …, n_r objects are indistinct from each other is

$$\frac{n!}{n_1! \cdots n_r!} \text{ where } n_1 + \cdots + n_r = n.$$

1.1.3 Combinations

Definition 1.1.2. (**Combination**) A combination of a set S is an unordered sequence (a set) $\gamma \subseteq S$.

Theorem 1.1.7. Suppose $0 \leq k \leq n$, the number of k element combinations of a set of n elements is

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

Proof. Let ${}_nC_k$ denote the number of k element combinations of a set of n elements.

Note that to produce a permutation, we perform the two tasks in sequence:

- T_1: Select a combination $\gamma \subseteq S$ containing k elements.
- T_2: Choose a particular permutation π of γ.

Note that there are $k!$ permutations of γ, so by the product principle, we have

$$\begin{aligned} {}_nC_k \cdot k! &= {}_nP_k = \frac{n!}{(n-k)!} \\ \iff {}_nC_k &= \frac{n!}{k!(n-k)!} \end{aligned}$$

□

Theorem 1.1.8. Suppose we have $k \geq 2$ blocks each with $n_1, \ldots, n_k$ elements such that $n_1 + \cdots + n_k = n$, the number of such blocks on a set of n elements is

$$\binom{n}{n_1, \ldots, n_k} = \frac{n!}{n_1! \cdots n_k!}.$$

Proof. Suppose we have a set of n elements and $k \geq 2$ blocks each with $n_1, \ldots, n_k$ elements such that $n_1 + \cdots + n_k = n$.

We have k tasks in sequence:

- T_1: choose a combination of n_1 elements from n elements for the 1st block,
- T_2: choose a combination of n_2 elements from $n - n_1$ elements for the 2nd block,
- $\vdots$
- T_k: choose a combination of n_k elements from $n - n_1 - \cdots - n_{k-1}$ for the kth block.

So by the product principle of counting

$$\begin{aligned}\binom{n}{n_1, \ldots, n_k} &= \binom{n}{n_1} \cdot \binom{n-n_1}{n_2} \cdots \binom{n-n_1-\cdots-n_{k-1}}{n_k} \\ &= \frac{n!}{n_1!(n-n_1!)} \frac{(n-n_1)!}{n_2!(n-n_1-n_2)!} \cdots \frac{(n-n_1-\cdots-n_{k-1}-n_k)!}{n_k!(n-n_1-\cdots-n_{k-1}-n_k)!} \\ &= \frac{n!}{n_1!n_2!\cdots n_k!}\end{aligned}$$

□

1.2 Axioms of Probability

- A random experiment has *outcomes*, *events* and *probability*.
- The probability of some event ω occurring is denoted by $P(\omega)$.

Definition 1.2.1. (**Probability Space**) A probability space is a triple $(\Omega, \mathcal{F}, P)$, where Ω is the set of possible outcomes, referred to as the *sample space*. $\mathcal{F} \subseteq \mathcal{P}(\Omega)$ is the set of possible events. P is a probability measure of some event ω.

- Suppose ω_1 and ω_2 are events in the space $(\Omega, \mathcal{F}, P)$, then
  - **Union**: $\omega_1 \cup \omega_2$ is the event containing all outcomes of ω_1 **or** ω_2.
  - **Intersection**: $\omega_1 \cup \omega_2$ is the event containing all outcomes of ω_1 **and** ω_2.
  - **Complement**: $\overline{\omega}$ is the event containing all outcomes in Ω **not** in ω.
- Standard set theoretic laws holds (see discrete mathematics notes).

Definition 1.2.2. (**Frequentist Definition of Probability**) The probability measure P of some event ω is

$$P(\omega) = \lim_{n \to \infty} \frac{n(\omega)}{n},$$

where $n(E)$ is the number of trials where ω occurs and n is the number of trials.

1.2.1 Kolmogrov's Axioms

1. For all events $\omega \in \mathcal{F}$, the probability of ω occurs is in the range of 0 to 1, that is
$$\forall \omega \in \mathcal{F}. P(\omega) \in [0,1].$$

2. The probability of sample space Ω is $P(\Omega) = 1$.

3. For any collection of mutually exclusive (pairwise disjoint) outcomes $\omega_1, \ldots, \omega_n$ occurring satisfies:
$$P\left(\bigcup_{i=1}^{n} \omega_i\right) = \sum_{i=1}^{n} P(\omega_i).$$

1.2.2 Theorems

Theorem 1.2.1. (**Probability of the Empty Set**)
$$P(\emptyset) = 0.$$

Proof. Let $\omega_1 = A$ and $\omega_2 = B \setminus A$ where $A \subseteq B$ and $\omega_3 = \emptyset$. Note that we have $B = \omega_1 \cup \omega_2 \cup \omega_3$ and $\omega_1, \ldots, \omega_3$ are mutually exclusive. So by Kolmogrov's third axiom,
$$\begin{aligned} P(B) &= \sum_{i=1}^{3} P(\omega_i) \\ &= P(A) + P(B \setminus A) + P(\emptyset) \\ &= P(B) + P(\emptyset) \end{aligned}$$

Hence
$$P(\emptyset) = 0.$$

$\square$

Corollary 1.2.1.1. For all events $A, B \in \mathcal{F}$, if $A \subseteq B$ then $P(A) \leq P(B)$.

Theorem 1.2.2. (**Addition Law**) For all events $A, B \in \mathcal{F}$, the probability of A or B occurring is
$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Proof. Let $A, B \in \mathcal{F}$ be arbitrary events. Let us note that $A \setminus B$, $A \cap B$ and $B \setminus A$ are mutually exclusive, hence by Kolmogrov's 3rd axiom,

$$\begin{aligned} P(A) &= P(A \setminus B) + P(A \cap B) \\ P(B) &= P(B \setminus A) + P(A \cap B) \\ P(A \cup B) &= P(A \setminus B) + P(B \setminus A) + P(A \cap B) \end{aligned}$$

Hence

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

□

Corollary 1.2.2.1. (**Inclusion-Exclusion Principle**) For all events $\omega_1, \ldots, \omega_n \in \mathcal{F}$, we have

$$P\left(\bigcup_{i=1}^{n} \omega_i\right) = \sum_{i=1}^{n} (-1)^{i-1} \sum_{1 \le j_1 < \cdots < j_i \le n} P\left(\omega_{j_1} \cap \cdots \cap \omega_{j_i}\right).$$

Proof. Induction on n with a basis of $n = 2$. □

Corollary 1.2.2.2. (**Boole's Inequality**) For all events $\omega_1, \ldots, \omega_n$, we have

$$P\left(\bigcup_{i=1}^{n} \omega_i\right) \le \sum_{i=1}^{n} P(\omega_i).$$

Theorem 1.2.3. (**Complement Law**) For all events $A \in \mathcal{F}$, the probability of $\overline{A}$ occurring is

$$P(\overline{A}) = 1 - P(A).$$

Proof. By Kolmogrov's 1st axiom, we have $P(\Omega) = 1$. We note that for all events $A \in \mathcal{F}$, $\Omega = A \cup (\Omega \setminus A) = A \cup \overline{A}$. Hence

$$P(A \cup \overline{A}) = 1.$$

Note that A and $\overline{A}$ are mutually exclusive events, so by Kolmogrov's 3rd axiom,

$$\begin{aligned} P(A) + P(\overline{A}) &= 1 \\ \iff P(\overline{A}) &= 1 - P(A) \end{aligned}$$

□

Theorem 1.2.4. (**Subset Law**) For all events $A, B \in \mathcal{F}$, if $A \subseteq B$ the $P(A) \leq P(B)$.

Proof. Let $A, B \in \mathcal{F}$ be arbitrary events. Let us assume that $A \subseteq B$. Since A and $B \setminus A$ are mutually exclusive events, then by Kolmogrov's 3rd axiom, we have

$$\begin{aligned} P(B) &= P(A) + P(B \setminus A) \\ &\geq P(A) && \text{1st axiom. } P(B \setminus A) \geq 0 \end{aligned}$$

□

1.3 Conditional Probability

Definition 1.3.1. (**Conditional Probability**) Consider a probability space $(\Omega, \mathcal{F}, P)$. The conditional probability of event A given B has occurred (denoted $P(A \mid B)$) where $P(B) > 0$ is

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

- Conditional probability is equivalent to considering a probability space $(B, \mathcal{F}, P(\cdot \mid B))$ and considering the event $A \cap B$.

Theorem 1.3.1. (**Generalised Chain Rule**) For all events $A_1, \ldots, A_n \in \mathcal{F}$

$$P\left(\bigcap_{i=1}^{n} A_i\right) = P(A_1)P(A_2 \mid A_1) \cdots P(A_n \mid A_1 \cap \cdots \cap A_{n-1}).$$

1.3.1 Bayes' Theorem

Theorem 1.3.2. (**Bayes' Theorem**) For all events $A, B \in \mathcal{F}$ where $P(A), P(B) > 0$,

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

- A and B are referred as hypothesis and evidence respectively.

- $P(A)$ is the prior probability of the hypothesis.

- $P(B \mid A)$ is the likelihood.
- $P(B)$ is the "normalization" constant (ensures Kolmogrov's 1st axiom holds).
- $P(A \mid B)$ is the posterior probability of the hypothesis

Theorem 1.3.3. (**Partition Theorem**) For disjoint events $B_1, \dots, B_n$ such that $\bigcup B_i = \Omega$ (a partition of Ω), for all events $A \in \mathcal{F}$ we have

$$P(A) = \sum_{i=1}^{n} P(A \mid B_i) P(B_i).$$

Corollary 1.3.3.1. (**Bayes' Second Theorem**) For a partition $C_1, \dots, C_n$ on Ω, for all events $A, B \in \mathcal{F}$ where $P(A), P(B) > 0$,

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} = \frac{P(B \mid A)P(A)}{\sum_{i=1}^{n} P(B \mid C_i)P(C_i)}.$$

1.3.2 Independence

Definition 1.3.2. (**Independence**) For all events $A, B \in \mathcal{F}$, A and B are said to be independent if and only if

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

In general, the set of events $\mathcal{T} = \{A_1, \dots, A_n\} \subseteq \mathcal{F}$ are said to be independent if and only if

$$\forall \mathcal{S} \subseteq \mathcal{T}. P\left(\bigcap_{A_i \in \mathcal{S}} A_i\right) = \prod_{A_i \in \mathcal{S}} P(A_i).$$

Theorem 1.3.4. For all events $A, B \in \mathcal{F}$, A and B are independent if and only if

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

Proof. Let events $A, B \in \mathcal{F}$ be arbitrary. By definition, A and B are independent if and only if

$$\begin{aligned} P(A \cap B) &= P(A) \cdot P(B) \\ \iff P(A \mid B) &= \frac{P(A) \cdot P(B)}{P(B)} \\ &= P(A) \end{aligned}$$

□

Theorem 1.3.5. (**Independence Of Complements**) For all events $A, B \in \mathcal{F}$, if A and B are independent, then A and $\overline{B}$ are independent.

Proof. Let events $A, B \in \mathcal{F}$ be arbitrary. Let us assume that A and B are independent, that is to say

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

Consider $P(A \cap \overline{B})$. So we have

$$\begin{aligned} P(A \cap \overline{B}) &= P(A \setminus B) \\ &= P(A) - P(A \cap B) && \text{(see theorem ??)} \\ &= P(A) - P(A) \cdot P(B) \\ &= P(A) \cdot (1 - P(B)) \\ &= P(A) \cdot P(\overline{B}) && \text{(complement law)} \end{aligned}$$

Hence A and $\overline{B}$ are independent. □

Definition 1.3.3. (**Conditional Independence**) For all events $A, B, C \in \mathcal{F}$, A and B are said to be conditionally independent given C if and only if

$$P(A \cap B \mid C) = P(A \mid C) \cdot P(B \mid C).$$

Theorem 1.3.6. For all events $A, B, C \in \mathcal{F}$, A and B are said to be conditionally independent given C if and only if

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

2 Random Variables

- **Motivation**: desire to work with a real-valued function on probability space $(\Omega, \mathcal{F}, P)$ instead of outcomes $\omega \in \Omega$.
- This defines the notation of a *random variable*.

Definition 2.0.1. (**Random Variable**) A random variable on the probability space $(\Omega, \mathcal{F}, P)$ is a total function $X : \Omega \to \mathbb{R}$ s.t.

$$\forall x \in \mathbb{R}.\, \{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F}.$$

Definition 2.0.2. (**Cumulative Distribution Function**) For a random variable X on $(\Omega, \mathcal{F}, P)$, the cumulative distribution function (c.d.f) of X is defined as

$$F_X(x) = P(X \leq x) = P(\{\omega \in \Omega : X(\omega) \leq x\}) : \mathbb{R} \to [0, 1].$$

Theorem 2.0.1. (**Properties of c.d.f**) For random variable X on $(\Omega, \mathcal{F}, P)$, the c.d.f F_X satisfies

1. The direct image $\overrightarrow{F_X}(\mathbb{R}) = [0, 1]$.
2. If $x < y$ then $F_X(x) < F_X(y)$.
3. $\lim_{x \to -\infty} F_X(x) = 0$ and $\lim_{x \to \infty} F_X(x) = 1$.
4. If $a < b$, then
$$P(a < X \leq b) = F_X(b) - F_X(a).$$

2.1 Discrete Random Variables

Definition 2.1.1. (**Discrete Random Variable**) A random variable $X : \Omega \to \mathbb{R}$ on the probability space $(\Omega, \mathcal{F}, P)$ is discrete if if

1. The direct image $\overrightarrow{X}(\Omega) = \{X(\omega) \in \mathbb{R} : \omega \in \Omega\}$ is a countable set.

Definition 2.1.2. (**Probability Mass Function**) For a discrete random variable X on $(\Omega, \mathcal{F}, P)$, the probability mass function (p.m.f) of X, denoted $p_X : \mathbb{R} \to [0, 1]$ is defined as

$$P(X = x) = p_X(x) = \begin{cases} P(\{w \in \Omega : X(\omega) = x\}) & \text{if } x \in \overrightarrow{X}(\Omega) \\ 0 & \text{otherwise} \end{cases}.$$

- By Kolmogrov's axioms, the p.m.f p_X satisfies
  1. For all $x \in \overrightarrow{X}(\Omega)$, $p_X(x) \geq 0$.
  2. For any interval $\mathcal{I}$, $P(X \in \mathcal{I}) = \sum_{x \in \mathcal{I}} p_X(x)$
  3. $\sum_{x \in \overrightarrow{X}(\Omega)} p_X(x) = 1$.
- The p.m.f p_X describes a **distribution** of probabilities over the outcomes of X.

Definition 2.1.3. (**Cumulative Distribution Function**) For a discrete random variable X on $(\Omega, \mathcal{F}, P)$, the cumulative distribution of X is

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

2.1.1 Expectation and Variance

Definition 2.1.4. (**Expectation**) For a discrete random variable X on $(\Omega, \mathcal{F}, P)$, the expectation of X is defined as

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

provided the sum is absolutely convergent, that is

$$\sum_{x \in \overrightarrow{X}(\Omega)} |x P(X = x)| < \infty.$$

- $\mathbb{E}[X]$ is often referred to as the expected value, mean or the first moment.

- $\mathbb{E}[X^n]$ is the nth moment.

Theorem 2.1.1. For a discrete random variable X on $(\Omega, \mathcal{F}, P)$ and $g : \mathbb{R} \to \mathbb{R}$ is some transformation, then

$$\mathbb{E}[g(X)] = \sum_{x \in \overrightarrow{X}(\Omega)} g(x)P(X = x),$$

provided the sum is absolutely convergent.

Proof. Let X be a discrete random variable on $(\Omega, \mathcal{F}, P)$ and $g : \mathbb{R} \to \mathbb{R}$. Let Y be a discrete random variable s.t $Y(\omega) = g(X)(\omega) = g(X(\omega))$. Consider the expectation of Y, so

$$\begin{aligned}
\mathbb{E}[g(X)] = \mathbb{E}[Y] &= \sum_{y \in \overrightarrow{g}(\overrightarrow{X}(\Omega))} yP(Y = y) \\
&= \sum_{y \in \overrightarrow{g}(\overrightarrow{X}(\Omega))} y \sum_{x \in \overrightarrow{X}(\Omega) : g(x) = y} P(X = x) \\
&= \sum_{x \in \overrightarrow{X}(\Omega)} g(x)P(X = x)
\end{aligned}$$

□

Corollary 2.1.1.1. (**Linearity of Expectation**) For all $a, b \in \mathbb{R}$

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

Proof. Let X be a discrete random variable on $(\Omega, \mathcal{F}, P)$, and $a, b \in \mathbb{R}$ be arbitrary. Let $g(x) = ax + b : \mathbb{R} \to \mathbb{R}$. So

$$\begin{aligned}
\mathbb{E}[aX + b] &= \sum_{x \in \overrightarrow{X}(\Omega)} (ax + b)P(X = x) \\
&= a \sum_{x \in \overrightarrow{X}(\Omega)} xP(X = x) + b \sum_{x \in \overrightarrow{X}(\Omega)} P(X = x) \\
&= a\mathbb{E}[X] + b
\end{aligned}$$

□

Definition 2.1.5. (**Variance**) For the random variable X on $(\Omega, \mathcal{F}, P)$, the variance of X is

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

- Expectation defines the "central location" $\mathbb{E}[X]$.
- Variance is the expected deviation (dispersion) of X about it's expected value
- Deviation is $|X - \mathbb{E}[X]|$, but, $|\cdot|$ is difficult $\implies$ use $(X - \mathbb{E}[X])^2$
- The expected deviation in correct units is the **standard deviation**

$$\sigma(X) = \sqrt{\operatorname{Var}[X]} = \sqrt{\mathbb{E}\left[(X - \mathbb{E}[X])^2\right]}.$$

Theorem 2.1.2. For discrete random variable X on $(\Omega, \mathcal{F}, P)$, the variance of X is

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

Proof. Let X be a discrete random variable on $(\Omega, \mathcal{F}, P)$. Let $\mu = \mathbb{E}[X]$. So

$$\begin{aligned}
\operatorname{Var}[X] &= \mathbb{E}[(X - \mu)^2] \\
&= \sum_{x \in \overrightarrow{X}(\Omega)} (x - \mu)^2 P(X = x) \\
&= \sum_{x \in \overrightarrow{X}(\Omega)} (x^2 - 2x\mu + \mu^2) P(X = x) \\
&= \sum_{x \in \overrightarrow{X}(\Omega)} x^2 P(X = x) - 2\mu \left(\sum_{x \in \overrightarrow{X}(\Omega)} x P(X = x) \right) + \mu^2 \\
&= \mathbb{E}[X^2] - 2\mu^2 + \mu^2 = \mathbb{E}[X^2] - (\mathbb{E}[X])^2
\end{aligned}$$

□

Theorem 2.1.3. (**Non-linearity of Variance**) For a discrete random variable X on $(\Omega, \mathcal{F}, P)$. For all $a, b \in \mathbb{R}$

$$\operatorname{Var}[aX + b] = a^2 \operatorname{Var}[X].$$

Proof. Let X be a discrete random variable on $(\Omega, \mathcal{F}, P)$. Let $a, b \in \mathbb{R}$ be arbitrary.

$$\begin{aligned}
\operatorname{Var}[aX+b] &= \mathbb{E}\left[(aX + b - \mathbb{E}[aX+b])^2\right] \\
&= \mathbb{E}\left[(aX + b - a\mathbb{E}[X] - b)^2\right] \\
&= \mathbb{E}\left[a^2(X - \mathbb{E}[X])^2\right] \\
&= a^2 \operatorname{Var}[X]
\end{aligned}$$

□

2.1.2 Discrete Distributions

Bernoulli Distribution

Definition 2.1.6. (**Bernoulli Distribution**) For discrete random variable X on $(\Omega, \mathcal{F}, P)$, X has the Bernoulli distribution with parameter $0 \leq p \leq 1$ iff

1. $\overrightarrow{X}(\Omega) = \{0, 1\}$

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

- **Notation**: $X \sim \operatorname{Bern}(p)$.
- $\mathbb{E}[X] = p$.
- $\operatorname{Var}[X] = p(1-p)$.
- **Description**: A single experiment with probability p of success.

Discrete Uniform Distribution

Definition 2.1.7. (**Discrete Uniform Distribution**) For discrete random variable X on $(\Omega, \mathcal{F}, P)$, X has discrete uniform distribution with parameter n iff

1. $\overrightarrow{X}(\Omega) = \{1, \ldots, n\}$

2.
$$P(X = x) = p_X(x) = \begin{cases} \frac{1}{n} & \text{if } x \in \{1, \ldots, n\} \\ 0 & \text{otherwise} \end{cases}.$$

- **Notation**: $X \sim \mathrm{U}(n)$.
- $\mathbb{E}[X] = (n+1)/2$.
- $\mathrm{Var}[X] = (n^2 - 1)/12$.
- **Description**: A single experiment with all outcomes equally likely.

Binomial Distribution

Definition 2.1.8. (**Binomial Distribution**) For discrete random variable X on $(\Omega, \mathcal{F}, P)$, X has the Binomial distribution with parameters $n \geq 1$ and $p \in [0, 1]$ iff

1. $\overrightarrow{X}(\Omega) = \{0, 1, \ldots, n\}$.
2.
$$P(X = x) = p_X(x) = \binom{n}{k} p^k (1-p)^{n-k}.$$

- **Notation**: $X \sim B(n, p)$.
- $\mathbb{E}[X] = np$.
- $\mathbb{E}[X] = np(1-p)$.
- **Description**: n independent trials with probability p of success.

Negative Binomial Distribution

Definition 2.1.9. (**Negative Binomial Distribution**) For discrete random variable X on $(\Omega, \mathcal{F}, P)$, X has the Negative Binomial distribution with parameters $r > 0$ and $p \in [0, 1]$ iff

1. $\overrightarrow{X}(\Omega) = \{r, r+1, \ldots, \}$
2.
$$P(X = x) = p_X(x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}.$$

- **Notation**: $X \sim \text{NB}(r, p)$
- $\mathbb{E}[X] = r/p$.
- $\text{Var}[X] = r(1-p)/p^2$
- **Description**: X models the number of independent trials until r successes.

Poisson Distribution

Definition 2.1.10. (**Poisson Dsitribution**) For discrete random variable X on $(\Omega, \mathcal{F}, P)$, X has the Poisson distribution with parameter $\lambda > 0$ iff

- $\overrightarrow{X}(\Omega) = \mathbb{N}$.
- $$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}.$$
- **Notation**: $X \sim \text{Poisson}(\lambda)$
- $\mathbb{E}[X] = \lambda$
- $\text{Var}[X] = \lambda$
- **Description**: X models number of successes over experiment duration, where λ is the rate of success.

Theorem 2.1.4. (**Binomial Approximated by Poisson Distribution**) Let X be a discrete random variable on $(\Omega, \mathcal{F}, P)$ with the binomial distribution with parameters n and p. Suppose n is "large" and p is "small", then X can be approximated by a Poisson distribution with parameter $\lambda = np$.

Proof. Let X be as described. Let $np = \lambda$. So

$$\begin{aligned}
P(X = x) &= \binom{n}{x} p^x (1-p)^{n-x} \\
&= \frac{n!}{(n-x)!x!} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} \\
&= \frac{n(n-1)\cdots(n-x+1)}{x!} \frac{\lambda^x}{n^x} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-x} \\
&= \frac{\lambda^x}{x!} \left(1 - \frac{\lambda}{n}\right)^n \frac{n(n-1)\cdots(n-x+1)}{n^x} \frac{1}{\left(1 - \frac{\lambda}{n}\right)^x}
\end{aligned}$$

Taking the limit as $n \to \infty$ yields

$$\begin{aligned}
\lim_{n\to\infty} P(X = x) &= \lim_{n\to\infty} \frac{\lambda^x}{x!} \left(1 - \frac{\lambda}{n}\right)^n \frac{n(n-1)\cdots(n-x+1)}{n^x} \frac{1}{\left(1 - \frac{\lambda}{n}\right)^x} \\
&= \frac{\lambda^x}{x!} e^{-\lambda}
\end{aligned}$$

Hence the result. □

Geometric Distribution

Definition 2.1.11. (**Geometric Distribution**) For discrete random variable X on $(\Omega, \mathcal{F}, P)$, X has the Geometric distribution with parameter $p \in [0, 1]$ iff

1. $\overrightarrow{X}(\Omega) = \mathbb{Z}^+$
2. $$P(X = x) = p_X(x) = p(1-p)^{x-1}.$$

- **Notation**: $X \sim \text{Geo}(p)$.
- $\mathbb{E}[X] = 1/p$.
- $\text{Var}[X] = (1-p)/p^2$.
- **Description**: X models the number of independent trials until first success with probability p.
- $\text{Geo}(p) \equiv \text{NB}(1, p)$

Hypergeometric Distribution

Definition 2.1.12. (**Hypergeometric Distribution**) For discrete random variable X on $(\Omega, \mathcal{F}, P)$, X has a Hypergeometric distribution with parameters $N \geq 0, 0 \leq m \leq N, 0 \leq n \leq N$ iff

1. $\overrightarrow{X}(\Omega) = \{\max(0, n - (N - m)), \ldots, \min(m, n)\}$
2. $$P(X = x) = p_X(x) = \frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}}.$$

- **Notation**: $X \sim \text{Hyp}(N, n, m)$.
- $\mathbb{E}[X] = n\frac{m}{N}$.
- $$\text{Var}[X] = n\frac{m}{N}\left(1 - \frac{m}{N}\right)\left(1 - \frac{n-1}{N-1}\right).$$
- **Description**: X models number of successes of sampling member with a feature without replacement in a sample size of n from a population of size N with m items with the feature

2.2 Continuous Random Variables

Definition 2.2.1. (**Continuous Random Variable**) A random variable $X : \Omega \to \mathbb{R}$ on $(\Omega, \mathcal{F}, P)$ is continuous iff there exists a function f_X s.t. for all intervals $\mathcal{I}$
$$P(X \in \mathcal{I}) = \int_{\mathcal{I}} f_X(x)\,\mathrm{d}x,$$
where f_X is the **probability density function** (p.d.f) of X.

- By Kolmogrov's Axioms, $\int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x = 1$.
- and,

$$\begin{aligned} P(X = x) &= 0 \\ P(X \leq x) &= P(X < x) \\ P(x \leq X \leq x + \mathrm{d}x) &= f_X(x)\,\mathrm{d}x \\ \forall x \in \mathbb{R}. f_X(x) &\geq 0 \end{aligned}$$

Definition 2.2.2. (**Cumulative Distribution Function**) For a continuous random variable X on $(\Omega, \mathcal{F}, P)$, the cumulative distribution of X is

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\,\mathrm{d}u\,.$$

- By the Fundamental Theorem of Calculus,

$$f_X(x) = \begin{cases} \frac{\mathrm{d}}{\mathrm{d}x} F_X(x) & \text{if the derivative exists at } x \\ 0 & \text{otherwise} \end{cases}.$$

2.2.1 Expectation and Variance

Definition 2.2.3. (**Expectation**) For a continuous random variable X on $(\Omega, \mathcal{F}, P)$, the expectation of X is defined as

$$\mathbb{E}[X] = \int_{x \in \overrightarrow{X}(\Omega)} x f_X(x)\,\mathrm{d}x\,.$$

provided the integral is absolutely convergent, that is

$$\int_{x \in \overrightarrow{X}(\Omega)} |x f_X(x)|\,\mathrm{d}x < \infty.$$

Theorem 2.2.1. For a continuous variable X on $(\Omega, \mathcal{F}, P)$ and $g : \mathbb{R} \to \mathbb{R}$ is some transformation, then

$$\mathbb{E}[g(X)] = \int_{x \in \overrightarrow{X}(\Omega)} g(x) f_X(x)\,\mathrm{d}x,$$

provided the integral is absolutely convergent.

Proof. Let X and g be as described. Let Y be a continuous random variable s.t $Y(\omega) = g(X)(\omega) = g(X(\omega))$. By inverse rule,

$$\frac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y) = \frac{1}{g'(g^{-1}(y))}.$$

Since $x = g^{-1}(y)$, it follows that

$$\mathrm{d}x = \frac{1}{g(g^{-1}(y))}\,\mathrm{d}y\,.$$

Similarly, note that

$$\begin{aligned} F_Y(y) &= P(g(X) \leq y) \\ &= P(X \leq g^{-1}(y)) \\ &= F_X(g^{-1}(y)) \end{aligned}$$

Hence by the chain rule,

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y} F_Y(y) = f_X(g^{-1}(y)) \frac{1}{g'(g^{-1}(y))}.$$

Consider the expectation of Y. So

$$\begin{aligned} \mathbb{E}[g(X)] = \mathbb{E}[Y] &= \int_{y \in \overrightarrow{Y}(\Omega)} y f_Y(y) \,\mathrm{d}y \\ &= \int_{y \in \overrightarrow{Y}(\Omega)} y f_X(g^{-1}(y)) \frac{1}{g'(g^{-1}(y))} \,\mathrm{d}y \\ &= \int_{x \in \overrightarrow{X}(\Omega)} g(x) f_X(x) \frac{1}{g'(x)} \left[g'(x) \,\mathrm{d}x\right] \qquad y = g(x) \text{ sub} \\ &= \int_{x \in \overrightarrow{X}(\Omega)} g(x) f_X(x) \,\mathrm{d}x \end{aligned}$$

□

Definition 2.2.4. (**Variance**) For the random variable X on $(\Omega, \mathcal{F}, P)$, the variance of X is

$$\mathrm{Var}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \int_{x \in \overrightarrow{X}(\Omega)} (x - \mathbb{E}[X])^2 f_X(x) \,\mathrm{d}x \,.$$

Theorem 2.2.2. For continuous random variable X on $(\Omega, \mathcal{F}, P)$, the variance of X is

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

Proof. Let X be a continuous random variable on $(\Omega, \mathcal{F}, P)$. Let $\mu = \mathbb{E}[X]$.

So

$$\begin{aligned}
\mathrm{Var}[X] &= \mathbb{E}[(X-\mu)^2] \\
&= \int_{x \in \overrightarrow{X}(\Omega)} (x-\mu)^2 f_X(x)\,\mathrm{d}x \\
&= \int_{x \in \overrightarrow{X}(\Omega)} (x^2 - 2\mu x + \mu^2) f_X(x)\,\mathrm{d}x \\
&= \int_{x \in \overrightarrow{X}(\Omega)} x^2 f_X(x)\,\mathrm{d}x - 2\mu \int_{x \in \overrightarrow{X}(\Omega)} x f_X(x)\,\mathrm{d}x + \mu^2 \\
&= \mathbb{E}[X^2] - 2\mu^2 + \mu^2 = \mathbb{E}[X^2] - (\mathbb{E}[X])^2
\end{aligned}$$

□

2.2.2 Continuous Distributions

Continuous Uniform Distribution

Definition 2.2.5. (**Continuous Uniform Distribution**) For continuous random variable X on $(\Omega, \mathcal{F}, P)$, X has continuous uniform distribution with parameters $\alpha < \beta$ iff

1. $\overrightarrow{X}(\Omega) = [a, b]$
2. $$f_X(x) = \begin{cases} \frac{1}{\beta - \alpha} & \alpha \leq x \leq \beta \\ 0 & \text{otherwise} \end{cases}.$$

- **Notation**: $X \sim \mathrm{U}[\alpha, \beta]$.
- $\mathbb{E}[X] = (\alpha + \beta)/2$.
- $\mathrm{Var}[X] = (\beta - \alpha)^2/12$.
- **Description**: A single experiment with all outcomes equally likely.

Exponential Distribution

Definition 2.2.6. (**Exponetial Distribution**) For continuous random variable X on $(\Omega, \mathcal{F}, P)$, X has the Exponential distribution with parameter λ iff

1. $\overrightarrow{X}(\Omega) = \mathbb{R}_{\geq 0}$.
2. $$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}.$$

- **Notation**: $X \sim \text{Exp}(\lambda)$.
- $\mathbb{E}[X] = 1/\lambda$.
- $\text{Var}[X] = 1/\lambda^2$.
- **Description**: X models the time until an event (first success) occurs with rate λ.

Normal Distribution

Definition 2.2.7. (**Normal Distribution**) For continuous random variable X on $(\Omega, \mathcal{F}, P)$, X has the Normal (Gaussian) distribution with parameters $\mu \in \mathbb{R}, \sigma^2 > 0$ iff

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

- **Notation**: $X \sim \mathcal{N}(\mu, \sigma^2)$.
- $\mathbb{E}[X] = \mu$.
- $\text{Var}[X] = \sigma^2$.

Definition 2.2.8. (**Standard Normal Distribution**) The standard Gaussian distribution is $\mathcal{N}(0, 1)$.

- Transformation from $X \sim \mathcal{N}(\mu, \sigma^2)$ using
$$Z = \frac{X - \mu}{\sigma},$$
so $Z \sim \mathcal{N}(0, 1)$ by linearity of expectation and non-linearity of variance. So
$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$
- The c.d.f of Z, $F_Z(z)$ is denoted by $\Phi(z)$, with $\Phi(-z) = 1 - \Phi(z)$.

2.3 Independence, Joint and Conditional Distributions

2.3.1 Joint Distributions

Definition 2.3.1. (**Joint Cumulative Distribution**) For two random variables X, Y on $(\Omega, \mathcal{F}, P)$, the join cumulative distribution F_{XY} is

$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y) = P\left(\{\omega \in \Omega : X(\omega) \leq x\} \cap \{\omega \in \Omega : Y(\omega) \leq y\}\right).$$

Discrete Case

Definition 2.3.2. (**Joint Probability Mass Function**) For two discrete random variables X, Y on $(\Omega, \mathcal{F}, P)$, the joint probability mass function is

$$\begin{aligned} P(X = x, Y = y) &= p_{X,Y}(x, y) \\ &= \begin{cases} P\left(\{\omega \in \Omega : X(\omega) = x \wedge Y(\omega) = y\}\right) & \text{if } x \in \overrightarrow{X}(\Omega), y \in \overrightarrow{Y}(\Omega) \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

- By Kolmogorov's axioms, must satisfy

$$\sum_{x \in \overrightarrow{X}(\Omega), y \in \overrightarrow{Y}(\Omega)} p_{X,Y}(x, y) = 1.$$

- So for all domains $\mathcal{D}$

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

Definition 2.3.3. (**Marginal Probability Mass Functions**) For two discrete random variables X, Y on $(\Omega, \mathcal{F}, P)$. Let $p_{X,Y}$ be the joint p.m.f of X and Y. Then p_X and p_Y are marginal probability mass functions of X and Y, where

$$\begin{aligned} p_X(x) &= \sum_{y \in \overrightarrow{Y}(\Omega)} p_{X,Y}(x, y) \\ p_Y(y) &= \sum_{x \in \overrightarrow{X}(\Omega)} p_{X,Y}(x, y) \end{aligned}$$

Definition 2.3.4. (**Joint Expectation**) For two discrete random variables X, Y on $(\Omega, \mathcal{F}, P)$ with joint p.m.f $p_{X,Y}$ and function $g : \mathbb{R}^2 \to \mathbb{R}$, then

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

Theorem 2.3.1. (**Linearity and Monotone**)

1. For two discrete random variables $X \leq Y$ on $(\Omega, \mathcal{F}, P)$, $\mathbb{E}[X] \leq \mathbb{E}[Y]$
2. For arbitrary discrete random variables $X_1, \dots, X_n$ on $(\Omega, \mathcal{F}, P)$,
$$\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i].$$

Continuous Case

Definition 2.3.5. (**Jointly Continuous**) For tow continuous random variables X, Y on $(\Omega, \mathcal{F}, P)$, X, Y are said to be jointly continuous if there exists a joint density function $f_{X,Y}$ s.t for all domains $\mathcal{D}$

$$P((X,Y) \in \mathcal{D}) = \iint_{\mathcal{D}} f_{X,Y}(x,y) \,\mathrm{d}x \,\mathrm{d}y \,.$$

- By Kolmogorov's axioms, must satisfy
$$\iint_{\overrightarrow{X}(\Omega) \times \overrightarrow{Y}(\Omega)} f_{X,Y}(x,y) \,\mathrm{d}x \,\mathrm{d}y = 1.$$
- The joint c.d.f is given by
$$f_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y).$$

Definition 2.3.6. (**Marginal Probability Density Functions**) For two jointly continuous random variables X, Y on $(\Omega, \mathcal{F}, P)$ with joint density function $f_{X,Y}(x,y)$. The p.d.fs f_X and f_Y are the marginal probability density functions of X and Y where

$$f_X(x) = \int_{y \in \overrightarrow{Y}(\Omega)} f_{X,Y}(x,y) \,\mathrm{d}y$$
$$f_Y(x) = \int_{x \in \overrightarrow{X}(\Omega)} f_{X,Y}(x,y) \,\mathrm{d}x$$

Definition 2.3.7. (**Joint Expectation**) For two discrete random variables X, Y on $(\Omega, \mathcal{F}, P)$ with joint p.d.f $f_{X,Y}$ and function $g : \mathbb{R}^2 \to \mathbb{R}$, then

$$\mathbb{E}[g(X,Y)] = \iint_{\overrightarrow{X}(\Omega)\times\overrightarrow{Y}(\Omega)} g(x,y) f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y\,.$$

Theorem 2.3.2. (**Linearity and Monotone**)

1. For two continuous random variables $X \leq Y$ on $(\Omega, \mathcal{F}, P)$, $\mathbb{E}[X] \leq \mathbb{E}[Y]$
2. For arbitrary continuous random variables $X_1, \ldots, X_n$ on $(\Omega, \mathcal{F}, P)$,

$$\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i].$$

Mixed Case

Definition 2.3.8. (**Mixed Joint Densitry Function**) For a continuous random variable X and a discrete random variable Y on $(\Omega, \mathcal{F}, P)$. The mixed joint density is defined as

$$f_X(x,y) = f_X(x \mid Y = y) \cdot P(Y = y) = P(Y = y \mid X = x) f_X(x)$$

2.3.2 Conditional Distributions

Definition 2.3.9. (**Conditional Probability Mass Function**) For two discrete random variables X, Y on $(\Omega, \mathcal{F}, P)$, the conditional probability mass function of X given $Y = y$ is defined as

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

Definition 2.3.10. (**Conditional Probability Density Function**) For two continuous random variables X, Y on $(\Omega, \mathcal{F}, P)$, the conditional probability density function of X given $Y = y$ is defined as

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

where $f_Y(y) > 0$.

- Conditional p.d.f not v. intuitive, consider

$$\begin{aligned} P(X = x + \mathrm{d}x \mid Y = y + \mathrm{d}y) &= \frac{P(X = x + \mathrm{d}x \mid Y = y + \mathrm{d}y)}{P(Y = y + \mathrm{d}y)} \\ &= \frac{f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y}{f_Y(y)\,\mathrm{d}y} \\ &= \frac{f_{X,Y}(x,y)}{f_Y(y)}\,\mathrm{d}x = f_X(x \mid Y = y)\,\mathrm{d}x \end{aligned}$$

Definition 2.3.11. (**Conditional Expectation**) For two random variables X, Y on $(\Omega, \mathcal{F}, P)$, the conditional expectation of X given $Y = y$ is given by

$$\begin{aligned} \mathbb{E}[X \mid Y = y] &= \sum_{x \in \overrightarrow{X}(\Omega)} xP(X = x \mid Y = y) && \text{(discrete case)} \\ &= \int_{x \in \overrightarrow{X}(\Omega)} x f_X(x \mid Y = y)\,\mathrm{d}x && \text{(continuous case)} \end{aligned}$$

Theorem 2.3.3. (**Law of Iterated Expectation**) For two random variables X, Y on $(\Omega, \mathcal{F}, P)$,

$$\mathbb{E}[X] = \mathbb{E}\left[\mathbb{E}[X \mid Y]\right].$$

Proof. For the discrete case, let X, Y be two discrete random variables on $(\Omega, \mathcal{F}, P)$. So

$$\begin{aligned} \mathbb{E}\left[[X \mid Y]\right] &= \sum_{y \in \overrightarrow{Y}(\Omega)} \mathbb{E}[X \mid Y = y]P(Y = y) \\ &= \sum_{y \in \overrightarrow{Y}(\Omega)} \left(\sum_{x \in \overrightarrow{X}(\Omega)} xP(X = x \mid Y = y) \right) P(Y = y) \\ &= \sum_{y \in \overrightarrow{Y}(\Omega)} \sum_{x \in \overrightarrow{X}(\Omega)} x \frac{P(X = x, Y = y)}{P(Y = y)} P(Y = y) \\ &= \sum_{y \in \overrightarrow{Y}(\Omega)} \sum_{x \in \overrightarrow{X}(\Omega)} xP(X = x, Y = y) \\ &= \mathbb{E}[X] \end{aligned}$$

For the continuous case, let X, Y be two continuous random variables on $(\Omega, \mathcal{F}, P)$. So

$$\begin{aligned}
\mathbb{E}\left[[X \mid Y]\right] &= \int_{y \in \overrightarrow{Y}(\Omega)} \mathbb{E}[X \mid Y = y] f_Y(y)\,\mathrm{d}y \\
&= \int_{y \in \overrightarrow{Y}(\Omega)} \left(\int_{x \in \overrightarrow{X}(\Omega)} x f_X(x \mid Y = y)\,\mathrm{d}x \right) f_Y(y)\,\mathrm{d}y \\
&= \iint_{(x,y) \in \overrightarrow{X}(\Omega) \times \overrightarrow{Y}(\Omega)} x \frac{f_{X,Y}(x,y)}{f_Y(y)} f_Y(y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \iint_{(x,y) \in \overrightarrow{X}(\Omega) \times \overrightarrow{Y}(\Omega)} x f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \mathbb{E}[X]
\end{aligned}$$

□

2.3.3 Independence

Definition 2.3.12. (**Independent Discrete Random Variables**) For two discrete random variables X, Y on $(\Omega, \mathcal{F}, P)$. X and Y are independent iff

$$P(X = x, Y = y) = P(X = x) \cdot P(Y = y).$$

Definition 2.3.13. (**Independent Continuous Random Variables**) For two continuous random variables X, Y on $(\Omega, \mathcal{F}, P)$. X and Y are independent iff

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

Theorem 2.3.4. (**Expectation of Independent Random Variables**) For two independent random variables X and Y and function $g, h : \mathbb{R} \to \mathbb{R}$,

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

Proof. For the discrete case, let X, Y be two independent discrete random variables on $(\Omega, \mathcal{F}, P)$. So

$$\begin{aligned}
\mathbb{E}[g(X) \cdot h(Y)] &= \sum_{x \in \overrightarrow{X}(\Omega), y \in \overrightarrow{Y}(\Omega)} g(x)h(y)P(X = x, Y = y) \\
&= \sum_{x \in \overrightarrow{X}(\Omega), y \in \overrightarrow{Y}(\Omega)} g(x)h(y)P(X = x)P(Y = y) \\
&= \left(\sum_{x \in \overrightarrow{X}(\Omega)} g(x)P(X = x) \right) \left(\sum_{y \in \overrightarrow{Y}(\Omega)} h(y)P(Y = y) \right) \\
&= \mathbb{E}[g(X)] \cdot \mathbb{E}[h(Y)]
\end{aligned}$$

For the continuous case, let X, Y be two independent continuous random variables on $(\Omega, \mathcal{F}, P)$. So

$$\begin{aligned}
\mathbb{E}[g(X) \cdot h(Y)] &= \iint_{\overrightarrow{X}(\Omega) \times \overrightarrow{Y}(\Omega)} g(x)h(y)f_{X,Y}(x, y)\, \mathrm{d}x\, \mathrm{d}y \\
&= \iint_{\overrightarrow{X}(\Omega) \times \overrightarrow{Y}(\Omega)} g(x)h(y)f_X(x)f_Y(y)\, \mathrm{d}x\, \mathrm{d}y \\
&= \int_{\overrightarrow{X}(\Omega)} g(x)f_X(x)\, \mathrm{d}x \cdot \int_{\overrightarrow{Y}(\Omega)} h(y)f_Y(y)\, \mathrm{d}y \\
&= \mathbb{E}[g(X)] \cdot \mathbb{E}[h(X)]
\end{aligned}$$

□

2.4 Covariance and Correlation

2.4.1 Covariance

Definition 2.4.1. (**Covariance**) For two random variables X, Y on $(\Omega, \mathcal{F}, P)$. The **covariance** of X and Y is

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

Theorem 2.4.1. For two random variables X, Y on $(\Omega, \mathcal{F}, P)$. The **covariance** of X and Y is

$$\mathrm{Cov}[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X] \cdot \mathbb{E}[Y].$$

Proof. Let X, Y be two arbitrary random variables on $(\Omega, \mathcal{F}, P)$. So we have

$$\begin{aligned}
\mathrm{Cov}[X,Y] &= \mathbb{E}\left[(X-\mathbb{E}[X])\cdot(Y-\mathbb{E}[Y])\right] \\
&= \mathbb{E}\left[XY - X\cdot\mathbb{E}[Y] - Y\cdot\mathbb{E}[X] + \mathbb{E}[X]\cdot\mathbb{E}[Y]\right] \\
&= \mathbb{E}[XY] - \mathbb{E}\left[\mathbb{E}[Y]\cdot X\right] - \mathbb{E}\left[\mathbb{E}[X]\cdot Y\right] + \mathbb{E}[X]\cdot\mathbb{E}[Y] \\
&= \mathbb{E}[XY] - \mathbb{E}[Y]\cdot\mathbb{E}\left[X\right] - \mathbb{E}[X]\cdot[Y] + \mathbb{E}[X]\cdot\mathbb{E}[Y] \\
&= \mathbb{E}[XY] - \mathbb{E}[X]\cdot\mathbb{E}[Y]
\end{aligned}$$

□

- $\mathrm{Cov}[X,X] = \mathrm{Var}[X]$.
- If X, Y are independent, $\mathrm{Cov}[X,Y] = 0$.

Theorem 2.4.2. (**Variance-Covariance formula**) For arbitrary random variables $X_1, \ldots, X_n$ on $(\Omega, \mathcal{F}, P)$,

$$\mathbb{E}\left[\left(\sum_{i=1}^{n} X_i\right)^2\right] = \sum_{i=1}^{n}\mathbb{E}[X^2] + \sum_{i\neq j}\mathbb{E}[X_i\cdot X_j] \tag{i}$$

$$\mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n}\mathrm{Var}[X_i] + \sum_{i\neq j}\mathrm{Cov}[X_i, X_j] \tag{ii}$$

Proof. Let $X_1, \ldots, X_n$ be arbitrary random variables on $(\Omega, \mathcal{F}, P)$. We note that

$$\begin{aligned}
\left(\sum_{i=1}^{n} X_i\right)^2 &= \sum_{i=1}^{n}\sum_{j=1}^{n} X_i X_j \\
&= \sum_{i=1}^{n} X_i^2 + \sum_{i\neq j} X_i X_j
\end{aligned}$$

Applying the linearity of expectation yields (i). (ii) follows from (i),

$$\begin{aligned}\mathrm{Var}\left[\sum_{i=1}^n X_i\right] &= \mathbb{E}\left[\left(\sum_{i=1}^n X_i - \mathbb{E}\left[\sum_{i=1}^n X_i\right]\right)^2\right] \\ &= \mathbb{E}\left[\left(\sum_{i=1}^n (X_i - \mathbb{E}[X_i])\right)^2\right] \\ &= \sum_{i=1}^n \mathrm{Var}[X_i] + \sum_{i\neq j} \mathbb{E}\left[(X_i - \mathbb{E}[X_i])\,(X_j - \mathbb{E}[X_j])\right] \quad \text{(by (i))}\end{aligned}$$

□

Theorem 2.4.3. (**Linearity of Variance for Independence Random Variables**) For independent random variables $X_1, \dots, X_n$ on $(\Omega, \mathcal{F}, P)$,

$$\mathrm{Var}\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n \mathrm{Var}[X_i].$$

2.4.2 Correlation

- Let $X = I_A$ and $Y = I_B$, for two events $A, B \in \mathcal{F}$. Then $\mathbb{E}[X] = P(A)$ and $\mathbb{E}[Y] = P(B)$, and $\mathbb{E}[XY] = P(A \cap B)$ then

$$\mathrm{Cov}[X,Y] = P(A \cap B) - P(A)P(B) = P(A)\left[P(B \mid A) - P(B)\right].$$

- If $\mathrm{Cov}[X,Y] > 0 \implies P(B \mid A) > P(B)$. Then A and B are positively correlated (vice versa).
- If $P(B \mid A) = P(B)$ then X and Y are uncorrelated.

Definition 2.4.2. (**Correlation Coefficent**) Let X and Y be two random variables on $(\Omega, \mathcal{F}, P)$. The correlation coefficient of X and Y is defined by

$$\rho(X,Y) = \frac{\mathrm{Cov}[X,Y]}{\sqrt{\mathrm{Var}[X] \cdot \mathrm{Var}[Y]}}.$$

If $\mathrm{Var}[X], \mathrm{Var}[Y] = 0$, then $\rho(X,Y) = 0$.

Theorem 2.4.4. For two random variables X, Y on $(\Omega, \mathcal{F}, P)$,

(i) The correlation coefficient is scaling invariant, for all $a, b \in \mathbb{R}$, $\rho(X, Y) = \rho(aX, bY)$.

(ii) $\rho(X, Y) \in [-1, 1]$.

Proof. Let X, Y be as described. For (i), let $a, b \in \mathbb{R}$ be arbitrary. We note that by the linearity of expectation

$$\begin{aligned}\operatorname{Cov}[aX, bY] &= \mathbb{E}[aXbY] - \mathbb{E}[aX] \cdot \mathbb{E}[bY] \\ &= ab\left(\mathbb{E}[XY] - \mathbb{E}[X] \cdot \mathbb{E}[Y]\right) = ab \operatorname{Cov}[X, Y]\end{aligned}$$

Hence, by the non-linearity of variance

$$\begin{aligned}\rho(aX, bY) &= \frac{\operatorname{Cov}[aX, bY]}{\sqrt{\operatorname{Var}[aX] \cdot \operatorname{Var}[bY]}} \\ &= \frac{ab \operatorname{Cov}[X, Y]}{\sqrt{a^2 \operatorname{Var}[X] \cdot b^2 \operatorname{Var}[Y]}} \\ &= \rho(X, Y)\end{aligned}$$

For (ii), $\lambda \in \mathbb{R}$ be arbitrary. Let $V = X - \mathbb{E}[X]$ and $W = Y - \mathbb{E}[Y]$. We note that

$$\begin{aligned}&\mathbb{E}[(V + \lambda W)^2] \geq 0 \\ \iff &\mathbb{E}\left[V^2 + 2\lambda VW + \lambda^2 W^2\right] \geq 0 \\ \iff &\mathbb{E}[W^2]\lambda^2 + 2\mathbb{E}[VW]\lambda + \mathbb{E}[V^2] \geq 0 \\ \iff &\operatorname{Var}[Y]\lambda^2 + 2\operatorname{Cov}[X, Y]\lambda + \operatorname{Var}[X] \geq 0\end{aligned}$$

Note that $\operatorname{Var}[Y] \geq 0$, hence convex quadratic function of λ. Hence the function ≥ 0 iff the discriminant $\Delta \leq 0$. So we have

$$\begin{aligned}&4\operatorname{Cov}^2[X, Y] - 4\operatorname{Var}[X] \cdot \operatorname{Var}[Y] \leq 0 \\ \iff &\rho^2(X, Y) = \frac{\operatorname{Cov}^2[X, Y]}{\operatorname{Var}[X] \cdot \operatorname{Var}[Y]} \leq 1\end{aligned}$$

Hence $\rho(X, Y) \in [-1, 1]$ $\square$

3 Moment and Limit Theorems

3.1 Markov, Chebyshev and Jensen Inequalities

Theorem 3.1.1. (**Markov Inequality**) For random variable $X \geq 0$ on $(\Omega, \mathcal{F}, P)$, for $a > 0$, we have

$$P(X \geq a) \leq \frac{1}{a}\mathbb{E}[X].$$

Proof. Let X and a be as described. The crucial observation

$$I_{\{X \geq a\}} \leq \frac{1}{a}X.$$

By theorem ??,

$$P(X \geq a) = \mathbb{E}\left[I_{\{X \geq a\}}\right] \leq \frac{1}{a}\mathbb{E}[X].$$

$\square$

Theorem 3.1.2. (**Chebyshev Inequality**) For random variable X on $(\Omega, \mathcal{F}, P)$. If $\mathbb{E}[X] = \mu$ and $\mathrm{Var}[X] = \sigma^2$ are both finite, then for all $k > 0$,

$$P\left(|X - \mu| \geq k\right) \leq \frac{\sigma^2}{k^2}.$$

Proof. Let X, μ, σ^2 and k be as described. We note that

$$P(|X - \mu| \geq k) = P((X - \mu)^2 \geq k^2).$$

By the Markov Inequality, we have

$$P\left((X - \mu)^2 \geq k^2\right) \leq \frac{1}{k^2}\mathbb{E}[(X - \mu)^2] = \frac{\sigma^2}{k^2}$$

$\square$

Theorem 3.1.3. (**Jensen's Inequality**) For random variable X on $(\Omega, \mathcal{F}, P)$. If g is convex, then

$$\mathbb{E}[g(X)] \geq g\left(\mathbb{E}[X]\right).$$

Proof. Let X and g be as described. Define the line y

$$y(x) - g(x_0) = g'(x_0)(x - x_0).$$

for $x_0 = \mathbb{E}[X]$. Note that y is tangent to g at x_0. Since g is convex, it follows that

$$\forall x \in \mathbb{R}. y(x) \leq g(x).$$

Hence $y(X) \leq g(X)$. By theorem ?? we have

$$\begin{aligned}\mathbb{E}[g(X)] \geq \mathbb{E}[y(X)] &= \mathbb{E}\left[g'(x_0)(X - x_0) + g(x_0)\right] \\ &= g'(x_0)\mathbb{E}[X - x_0] + g(x_0) \\ &= g'(x_0)\left(\mathbb{E}[X] - \mathbb{E}[X]\right) + g\left(\mathbb{E}[X]\right) = g\left(\mathbb{E}[X]\right)\end{aligned}$$

□

3.2 Weak Law of Large Numbers

- Weak law of large numbers formalizes the idea "if event A occurs w/ probability p, then repeating the trial a large number of times n, $n(A)/n \to p$"

Theorem 3.2.1. (**Weak Law of Large Numbers**) Let $X, X_1, \ldots, X_n$ be independent and identically distributed random variables on $(\Omega, \mathcal{F}, P)$ w/ finite expectation and variance, then $\overline{X_n}$ converges to $\mathbb{E}[X]$, that is for any $\epsilon > 0$,

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X]\right| \geq \epsilon\right) = 0.$$

Proof. Let $X, X_1, \ldots, X_n$ be as described. Let us define

$$\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

then

$$\mathbb{E}[X_n] = \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i]$$

Since $X, X_1, \ldots, X_n$ are identically distributed, then

$$\forall 1 \leq i \leq n . \mathbb{E}[X_i] = \mathbb{E}[X].$$

Hence

$$\mathbb{E}[\overline{X_n}] = \frac{n\mathbb{E}[X]}{n} = \mathbb{E}[X].$$

Similarly, since $X, X_1, \ldots, X_n$ are independent, then by theorem ??

$$\text{Var}[\overline{X_n}] = \frac{\text{Var}[X]}{n}.$$

Instantiating Chebyshev's inequality with $X = \overline{X_n}, k = \epsilon$ gives us

$$P\left(|\overline{X_n} - \mathbb{E}[X]| \geq \epsilon\right) \leq \frac{\text{Var}[X]}{n\epsilon^2}.$$

By Kolmogorov's axioms,

$$P\left(\left|\overline{X_n} - \mathbb{E}[X]\right| \geq \epsilon\right) \geq 0.$$

So

$$0 \leq P\left(\left|\overline{X_n} - \mathbb{E}[X]\right| \geq \epsilon\right) \leq \frac{\text{Var}[X]}{n\epsilon^2}.$$

We have

$$\lim_{n\to\infty} \frac{\text{Var}[X]}{n\varepsilon^2} = 0.$$

So by the Squeeze Theorem,

$$\lim_{n\to\infty} P\left(\left|\overline{X_n} - \mathbb{E}[X]\right| \geq \epsilon\right) = 0.$$

□

3.3 Moment Generating Functions

Definition 3.3.1. For a random variable X on $(\Omega, \mathcal{F}, P)$, the moment generating function of X, ϕ_X is defined as

$$\phi_X(t) = \mathbb{E}[e^{tX}],$$

for all t such that the expectation consists.

Theorem 3.3.1. (**Moment of Moment Generating Function**) For random variable X on $(\Omega, \mathcal{F}, P)$ with moment generating function ϕ_X, the nth moment is

$$\mathbb{E}[X^n] = \left.\frac{\mathrm{d}^n}{\mathrm{d}x^n}\right|_{t=0} \phi_X(t).$$

Proof. Let X, ϕ_X be as described. Consider nth derivative of ϕ_X. So we have

$$\begin{aligned}
\frac{\mathrm{d}^n}{\mathrm{d}t^n}\phi_X(t) &= \frac{\mathrm{d}^n}{\mathrm{d}t^n}\mathbb{E}[e^{tX}] \\
&= \frac{\mathrm{d}^n}{\mathrm{d}t^n}\mathbb{E}\left[\sum_{m=0}^{\infty} \frac{(tX)^m}{m!}\right] \\
&= \frac{\mathrm{d}^n}{\mathrm{d}t^n}\sum_{m=0}^{\infty}\mathbb{E}\left[\frac{t^m X^m}{m!}\right] \\
&= \sum_{m=0}^{\infty}\frac{\mathrm{d}^n}{\mathrm{d}t^n}\left(\frac{t^m}{m!}\right)\mathbb{E}[X^m] \\
&= \sum_{m=n}^{\infty}\frac{m\cdots(m-(n+1))t^{m-n}}{m!}\mathbb{E}[X^m] \\
&= \sum_{m=n}^{\infty}\frac{m!t^{m-n}}{m!(m-n)!}\mathbb{E}[X^m] \\
&= \mathbb{E}[X^n] + \sum_{m=n+1}^{\infty}\frac{m!t^{m-n}}{m!(m-n)!}\mathbb{E}[X^m]
\end{aligned}$$

Setting $t = 0$ gives us the result. □

Theorem 3.3.2. (**Moment Generating Function of Independent Random Variables**) Let $X_1, \ldots, X_n$ be independent random variables on $(\Omega, \mathcal{F}, P)$

with moment generating functions $\phi_{X_i}(t)$. For all $k_1, \dots, k_n \in \mathbb{R}$,

$$\phi_X(t) = \prod_{i=1}^{n} \phi_{X_i}(k_i t),$$

for all t where

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

Proof. Let $X_1, \dots, X_n$ and ϕ_{X_i} be as described. Let $k_1, \dots, k_n \in \mathbb{R}$ be arbitrary. Define

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

Hence

$$\begin{aligned}
\phi_X(t) = \mathbb{E}\left[e^{tX}\right] &= \mathbb{E}\left[\exp\left(t\sum_{i=1}^{n} k_i X_i\right)\right] \\
&= \mathbb{E}\left[\prod_{i=1}^{n} e^{tk_iX_i}\right] \\
&= \prod_{i=1}^{n} \mathbb{E}[e^{tk_iX_i}] && \text{(theorem ??)} \\
&= \prod_{i=1}^{n} \phi_{X_i}(k_i t)
\end{aligned}$$

□

Theorem 3.3.3. (**Linear Transformation of Moment Generating Function**) For random variable X on $(\Omega, \mathcal{F}, P)$ with moment generating function ϕ_X. For all $\alpha, \beta \in \mathbb{R}$,

$$\phi_Z(t) = e^{\beta t}\phi_X(\alpha t),$$

where $Z = \alpha X + \beta$.

Proof. Let X, ϕ_X, α, β and Z be as described. So

$$\begin{aligned}\phi_Z(t) &= \mathbb{E}[e^{tZ}] \\ &= \mathbb{E}\left[\exp\left(t\left(\alpha X + \beta\right)\right)\right] \\ &= \mathbb{E}[e^{t\alpha X}e^{\beta t}] \\ &= e^{\beta t}\mathbb{E}[e^{(\alpha t)X}] \\ &= e^{\beta t}\phi_X(\alpha t)\end{aligned}$$

□

3.4 Central Limit Theorem

Theorem 3.4.1. Let $X, X_1, \ldots, X_n$ be independent and identically distributed random variables on $(\Omega, \mathcal{F}, P)$ w/ finite expectation μ and variance σ^2.

$$\lim_{n\to\infty} P\left(\frac{S_n - \mu n}{\sqrt{n}\sigma} \leq x\right) = \Phi(x),$$

where

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

Proof. Let $X, X_1, \ldots, X_n$ be as described. Let us define

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

So we have

$$\begin{aligned}Z_n &= \sum_{i=1}^{n} \frac{X_i - \mu}{\sqrt{n}\sigma} \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma}\end{aligned}$$

Let us define

$$Y_i = \frac{X_i - \mu}{\sigma}.$$

Hence

$$Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i.$$

Let us assume that for all $1 \leq i \leq n$, the moment generating function of X_i, ϕ_{X_i} exists and is finite. Hence by theorem ??, the moment generating function of Y_i is given by

$$\phi_{Y_i}(t) = \exp\left(-\frac{\mu}{\sigma}\right)\phi_{X_i}\left(\frac{1}{\sigma}t\right)$$

Let us now consider the moment generating function of Z_n. Since $Y_1, \ldots, Y_n$ are independent identically distributed random variables, then by theorem ??

$$\begin{aligned}
\phi_{Z_n}(t) &= \prod_{i=1}^{n} \phi_{Y_i}\left(\frac{t}{\sqrt{n}}\right) \\
&= \left(\phi_Y\left(\frac{t}{\sqrt{n}}\right)\right)^n
\end{aligned}$$

Now let us define $\mathcal{L}(t) = \ln \phi_Y(t)$. Now consider $\mathcal{L}(0), \mathcal{L}'(0)$ and $\mathcal{L}''(0)$. By theorem ?? we have

$$\begin{aligned}
\mathcal{L}(0) &= \ln \phi_Y(0) = \ln 1 = 0 \\
\mathcal{L}'(0) &= \frac{\phi_Y'(0)}{\phi_Y(0)} = \phi_Y'(0) = \mathbb{E}[Y] = 0 \\
\mathcal{L}''(0) &= \frac{\phi_Y(0)\phi_Y''(0) - [\phi_Y'(0)]^2}{[\phi_Y(0)]^2} = \frac{1 \cdot \mathbb{E}[Y^2] - 0^2}{1^2} = \mathbb{E}[Y^2] = 1
\end{aligned}$$

We wish to show that $\lim_{n\to\infty} \phi_{Z_n}(t) = e^{t^2/2}$ (the m.g.f of $\mathcal{N}(0,1)$). So

$$\begin{aligned}
\lim_{n\to\infty} \phi_{Z_n}(t) = \left(\phi_Y\left(\frac{t}{\sqrt{n}}\right)\right)^n = e^{t^2/2} \\
\iff \lim_{n\to\infty} n \ln\left(\phi_Y\left(\frac{t}{\sqrt{n}}\right)\right) = t^2/2
\end{aligned}$$

Hence we compute

$$\begin{aligned}
\lim_{n\to\infty} \frac{\mathcal{L}(t/\sqrt{n})}{n^{-1}} &= \lim_{n\to\infty} \frac{-\mathcal{L}'(t/\sqrt{n})n^{-3/2}t}{-2n^{-2}} \\
&= \lim_{n\to\infty} \frac{\mathcal{L}'(t/\sqrt{n})t}{2n^{-1/2}} \\
&= \lim_{n\to\infty} \frac{-\mathcal{L}''(t/\sqrt{n})n^{-3/2}t^2}{-2n^{-3/2}} \\
&= \lim_{n\to\infty} \mathcal{L}''\left(\frac{t}{\sqrt{n}}\right)\frac{t^2}{2} \\
&= \mathcal{L}''(0)\frac{t^2}{2} = \frac{t^2}{2}
\end{aligned}$$

□

4 Applications and Statistics

4.1 Statistics and Estimators

4.1.1 Random Samples

Definition 4.1.1. (**Random Sample**) Let $X_1, \ldots, X_n$ be random variables on $(\Omega, \mathcal{F}, P)$ with c.d.fs $F_1, \ldots, F_n$. $\langle X_i \rangle$ from a **random sample** of size n if $X_1, \ldots, X_n$ are independent and $F_1 = \cdots = F_n$.

- $X_1, \ldots, X_n$ are independent and identically distributed (i.i.d).

Definition 4.1.2. (**Data Set**) A dataset of the sample $\langle X_i \rangle$ of size n is the set of realizations of the variables $x_1 = X_1(\omega), \ldots, x_n = X_n(\omega)$, denoted $\langle X_i(\omega) \rangle$.

Definition 4.1.3. (**Statistic**) A statistic of a sample $\langle X_i \rangle$ is a function $f : \mathbb{R}^n \to \mathbb{R}$.

Definition 4.1.4. (**Empirical Distribution**) For random sample $\langle X_i \rangle$ of size n with c.d.f F, the empirical distribution is

$$F_n(x) = \frac{n(X_i \leq x)}{n},$$

where $n(X_i \leq x)$ is number of realizations of $x_i \leq x$.

- By weak law of large numbers,

$$\lim_{n\to\infty} F_n(x) = F(x).$$

4.1.2 Estimators

- **Motivation**: Parameters of distribution F unknown, desire to estimate them based on random sample $\langle X_i \rangle$
- This defines notation of *estimator*.

Definition 4.1.5. (**Estimator**) For random sample $\langle X_i \rangle$ of size n with distribution F indexed by population parameter θ. The random variable

$$\hat{\theta} = \delta(X_1, \ldots, X_n),$$

where $\delta : \mathbb{R}^n \to \mathbb{R}$ is an **estimator** of θ. A particular realization of $\hat{\theta}$ is an **estimate** of θ.

Definition 4.1.6. (**Bias of Estimator**) For random sample $\langle X_i \rangle$ of size n with population parameter θ with estimator $\hat{\theta}$. The bias of $\hat{\theta}$ is

$$\text{Bias}\left[\hat{\theta}\right] = \mathbb{E}\left[\hat{\theta} - \theta\right].$$

Definition 4.1.7. (**Unbiased Estimator**) For estimator $\hat{\theta}$ of population parameter θ. $\hat{\theta}$ is said to be **unbiased** iff

$$\text{Bias}\left[\hat{\theta}\right] = 0.$$

Theorem 4.1.1. (**Unbiased Estimator for Expectation**) For random sample $\langle X_i \rangle$ of size n with distribution F with finite expectation μ. Then

$$\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

the **sample mean**, is an unbiased estimator for μ.

Proof. Let $\langle X_i \rangle$ and $\overline{X_n}$ be as described. We wish to show that $\text{Bias}[\overline{X_n}] = 0$. So

$$\begin{aligned}\mathbb{E}[\overline{X_n} - \mu] &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[X_i] - \mu \\ &= \frac{n\mu}{n} - \mu = 0\end{aligned}$$

□

Theorem 4.1.2. (**Variance of Sample Mean**) For random sample $\langle X_i \rangle$ of size n. The variance of the sample mean $\overline{X_n}$ is

$$\text{Var}[\overline{X_n}] = \frac{\sigma^2}{n}.$$

Proof. Let $\langle X_i \rangle$ and $\overline{X_n}$ be as defined. Since $X_1, \ldots, X_n$ are independently and identically distributed with finite variance σ^2, by theorem ?? we have

$$\text{Var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \text{Var}[X_i] = n\sigma^2$$

and by the non-linearity of variance, we have

$$\begin{aligned}
\text{Var}[\overline{X_n}] = \text{Var}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{X_i}{n}\right] &= \frac{1}{n^2}\text{Var}\left[\sum_{i=1}^{n} X_i\right] \\
&= \frac{\sigma^2}{n}
\end{aligned}$$

□

Theorem 4.1.3. (**Unbiased Estimator for Variance**) For random sample $\langle X_i \rangle$ of size n with distribution F with finite variances σ^2. Then

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

is an unbiased estimator for σ^2.

Proof. Let $\langle X_i \rangle$ and S_n be as described. We wish to show that $\text{Bias}[S_n] = 0$. So we have

$$\begin{aligned}
\mathbb{E}[S_n] &= \mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X_n})^2\right] \\
&= \mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}\left((X_i - \mu) - (\overline{X_n} - \mu)\right)^2\right] \\
&= \frac{1}{n-1}\mathbb{E}\left[\sum_{i=1}^{n}(X_i - \mu)^2 - 2(\overline{X_n} - \mu)\sum_{i=1}^{n}(X_i - \mu) + n(\overline{X_n} - \mu)^2\right]
\end{aligned}$$

We note that

$$\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)=\overline{X_n}-\mu$$

Hence

$$\begin{aligned}
\mathbb{E}[S_n] &= \frac{1}{n-1}\mathbb{E}\left[\sum_{i=1}^{n}(X_i-\mu)^2-2n(\overline{X_n}-\mu)^2+n(\overline{X_n}-\mu)^2\right] \\
&= \frac{1}{n-1}\left(\mathbb{E}\left[\sum_{i=1}^{n}(X_i-\mu)^2\right]-n\mathbb{E}\left[(\overline{X_n}-\mu)^2\right]\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}\mathbb{E}[(X_i-\mu)^2]-n\frac{\sigma^2}{n}\right) \\
&= \frac{1}{n-1}\left(n\sigma^2-\sigma^2\right)=\sigma^2
\end{aligned}$$

Hence $\mathbb{E}[S_n-\sigma^2]=0$. □

Definition 4.1.8. (**Mean Squared Error of Estimator**) For random sample $\langle X_i \rangle$ of size n with population parameter θ. The **mean squared error** of the estimator $\hat{\theta}$ is

$$\text{MSE}\left[\hat{\theta}\right]=\mathbb{E}\left[(\hat{\theta}-\theta)^2\right].$$

- MSE quantifies random error of an estimator, considering $\mathbb{E}\left[\left|\hat{\theta}-\theta\right|\right]$ (but $|\cdot|$ is difficult $\implies$ $(\cdot)^2$)
- MSE can compare estimators: $\hat{\theta}_1$ and $\hat{\theta}_2$ are estimators for population parameter θ, if $\text{MSE}\left[\hat{\theta}_1\right]<\text{MSE}\left[\hat{\theta}_2\right]$ then $\hat{\theta}_1$ is better than $\hat{\theta}_2$.

Theorem 4.1.4. For random sample $\langle X_i \rangle$ of size n with population parameter θ. The **mean squared error** of the estimator $\hat{\theta}$ is

$$\text{MSE}\left[\hat{\theta}\right]=\text{Bias}^2\left[\hat{\theta}\right]+\text{Var}\left[\hat{\theta}\right].$$

Proof. Let $\langle X_i \rangle$ and $\hat{\theta}$ be as described. We have

$$\begin{aligned}
\text{Bias}^2\left[\hat{\theta}\right] + \text{Var}\left[\hat{\theta}\right] &= \left(\mathbb{E}\left[\hat{\theta} - \theta\right]\right)^2 + \mathbb{E}\left[\hat{\theta}^2\right] - \left(\mathbb{E}\left[\hat{\theta}\right]\right)^2 \\
&= \left(\mathbb{E}\left[\hat{\theta}\right]\right)^2 - 2\theta\mathbb{E}\left[\hat{\theta}\right] + \theta^2 + \mathbb{E}\left[\hat{\theta}^2\right] - \left(\mathbb{E}\left[\hat{\theta}\right]\right)^2 \\
&= \mathbb{E}\left[\hat{\theta}^2 - 2\theta \cdot \hat{\theta} + \theta^2\right] \\
&= \text{MSE}\left[\hat{\theta}\right]
\end{aligned}$$

□

4.2 Testing Probability Distributions

- **Motivation**: Desire to estimate properties of distributions very quickly w/ large domains e.g. $\mathbb{Z}$.
- e.g. testing lottery number distribution, birthday problem, etc.

Definition 4.2.1. (**Formal Model**) For discrete random variable X on $(\Omega, \mathcal{F}, P)$ w/ p.m.f $p_X = (p_1, \dots, p_n)$ where n is finite, test whether

1. p_X is approximate to distribution w/ p.m.f p_Y?
2. What is $\max p_X$?
3. Are two distributions w/ p.m.fs p_X and p_Y independent?

4.2.1 Testing Distributions

Definition 4.2.2. (**Distance Between Distributions**) For discrete random variables X, Y on $(\Omega, \mathcal{F}, P)$ w/ p.m.fs p_X and p_Y. Then the

1. L_1-distance between p_X and p_Y is

$$\|p_X - p_Y\|_1 = \sum_{x \in \overrightarrow{X}(\Omega) \cap \overrightarrow{Y}(\Omega)} |p_X(x) - p_Y(x)| \in [0, 2].$$

2. L_2-distance between p_X and p_Y is

$$\|p_X - p_Y\|_2 = \sqrt{\sum_{x \in \overrightarrow{X}(\Omega) \cap \overrightarrow{Y}(\Omega)} (p_X(x) - p_Y(x))^2} \in [0, \sqrt{2}].$$

3. L_∞-distance between p_X and p_Y is

$$\|p_X - p_Y\|_\infty = \max_{x \in \overrightarrow{X}(\Omega) \cap \overrightarrow{Y}(\Omega)} |p_X(x) - p_Y(x)| \in [0, 1].$$

Example 4.2.1. (**Testing Uniformity**) We wish to find an **efficient** (sub-linear) tester s.t. for any discrete random variable X on $(\Omega, \mathcal{F}, P)$ w/ *p.m.f* $p_X = (p_1, \ldots, p_n)$ where n is finite, and accuracy $0 < \epsilon < 1$,

- If $X \sim U(n)$, them $P(\textbf{Accept}) \geq 2/3$
- If X is ϵ-far from $Y \sim U(n)$, that is $\|p_X - p_Y\|_1 \geq \epsilon$. then $P(\textbf{Reject}) \geq 2/3$.

Let X, Y, p_X, p_Y and ϵ be as described. In the case $X \sim U(n)$, then L_1-distance $\|p_X - p_Y\| = 0$. So consider L_1-distance, recall that

$$\|p_X - p_Y\| = \sum_{x=1}^{n} \left| p_X(x) - \frac{1}{n} \right|,$$

so requires $\Omega(n)$ queries of p_X (**not suitable**). So consider L_2-distance instead. Note that

$$\begin{aligned}
\|p_X - p_Y\|_2^2 &= \sum_{x=1}^{n} \left(p_X(x) - \frac{1}{n} \right)^2 \\
&= \sum_{x=1}^{n} p_X(x)^2 - \frac{2}{n} \sum_{x=1}^{n} p_X(x) + \frac{n}{n^2} \\
&= \|p_X\|_2^2 - \frac{1}{n}
\end{aligned}$$

Note that $p_X(x)^2$ is the probability of a collision (See Birthday Paradox).

- If X is (close to) uniform, then expected number of samples until first collision is $\sqrt{n}$.

- If X is far from uniform, minimum number of samples until first collision is 2.

So potential **sub-linear** method to estimate $\|p_X\|_2^2$. Method description:

1. Define random sample $\langle X_i \rangle$ of size r from X. Obtain data set.
2. Define $I_{i,j} = I_{\{X_i = X_j\}}$ (indicator for collision)
3. Define estimator $\hat{\theta}$ of $\|p_X\|_2^2$ s.t

$$\hat{\theta} = \frac{1}{\binom{r}{2}} \sum_{1 \le i < j \le r} I_{i,j}.$$

   Apply to data set, return estimate.

function ESTIMATE-$\|p_X\|_2^2(r)$
    $\langle X_i(\omega) \rangle \leftarrow$ **data set of sample** $\langle X_i \rangle$
    $A \leftarrow$ **array of size** n
    for $x_i \in \langle X_i(\omega) \rangle$ **do**
        $A[x_i]$++
    end for
    $\sum_{1 \le i < j \le r} I_{i,j} \leftarrow \sum_{k=1}^{n} \binom{A[k]}{2}$
    return $\frac{1}{\binom{r}{2}} \sum_{1 \le i < j \le r} I_{i,j}$
end function

- **Complexity**: Directly computing $\sum_{1 \le i < j \le r} I_{i,j}$ takes $\Theta\left(\binom{r}{2}\right) = \Theta(r^2)$. Using above method w/ array A takes $O(r)$ time w/ identity

$$\begin{aligned} \sum_{1 \le i < j \le r} I_{i,j} &= \sum_{1 \le i < j \le r} \sum_{k=1}^{n} I_{\{X_i = X_j = k\}} \\ &= \sum_{k=1}^{n} \sum_{1 \le i < j \le r} I_{\{X_i = X_j = k\}} \\ &= \sum_{k=1}^{n} \binom{A[k]}{2} \end{aligned}$$

We shall now analyze our estimator $\hat{\theta}$ for $\|p_k\|_2^2$.

Theorem 4.2.1. (**Analysis of $\hat{\theta}$**) For all $r \geq 36\sqrt{n}/\epsilon^2$, ESTIMATE-$\|p_X\|_2^2$ returns $\hat{\theta}$ s.t

$$P\left(\left|\hat{\theta} - \|p_X\|_2^2\right| \geq \epsilon \cdot \|p_X\|_2^2\right) \leq \frac{1}{3}.$$

Proof. Let us consider $\mathbb{E}[\hat{\theta}]$. We have

$$\begin{aligned}
\mathbb{E}\left[\hat{\theta}\right] &= \frac{1}{\binom{r}{2}} \sum_{1 \leq i < j \leq r} \mathbb{E}[I_{i,j}] \\
&= \frac{1}{\binom{r}{2}} \sum_{1 \leq i < j \leq r} \sum_{x=1}^{n} P(X_i = x) \cdot P(X_j = x) \\
&= \frac{1}{\binom{r}{2}} \sum_{1 \leq i < j \leq r} \sum_{x=1}^{n} p_X(x)^2 \\
&= \|p_X\|_2^2
\end{aligned}$$

Let us now consider $\mathrm{Var}[\hat{\theta}]$. Recall that

$$\begin{aligned}
\mathrm{Var}\left[\hat{\theta}\right] &= \mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}\left[\hat{\theta}\right]\right)^2\right] \\
&= \mathbb{E}\left[\left(\frac{1}{\binom{r}{2}} \sum_{1 \leq i < j \leq r} I_{i,j} - \mathbb{E}\left[\frac{1}{\binom{r}{2}} \sum_{1 \leq i < j \leq r} I_{i,j}\right]\right)^2\right] \\
&= \frac{1}{\binom{r}{2}} \mathbb{E}\left[\left(\sum_{1 \leq i < j \leq r} I_{i,j} - \mathbb{E}\left[I_{i,j}\right]\right)^2\right]
\end{aligned}$$

Let us define $Y_{i,j} = I_{i,j} - \mathbb{E}[I_{i,j}]$ and $Y_i = \sum_{j=i+1}^{r} Y_{i,j}$. We note that $\mathbb{E}[Y_{i,j}] = 0$. So by theorem ??, we have

$$\mathbb{E}\left[\left(\sum_{1 \leq i < j \leq r} Y_{i,j}\right)^2\right] = \underbrace{\sum_{1 \leq i < j \leq r} \mathbb{E}\left[Y_{i,j}^2\right]}_{A} + \underbrace{\sum_{i \neq j \neq k \neq \ell} \mathbb{E}[Y_{i,j} Y_{k,\ell}]}_{B} + 3! \cdot \underbrace{\sum_{1 \leq i < j < k \leq r} \mathbb{E}\left[Y_{i,j} Y_{i,k}\right]}_{C}$$

We note that

$$
\begin{aligned}
A &= \sum_{1\le i<j\le r} \mathbb{E}[Y_{i,j}^2] \le \sum_{1\le i<j\le r} \mathbb{E}[I_{i,j}^2] = \binom{r}{2}\|p_X\|_2^2 \\
B &= \sum_{i\ne j\ne k\ne \ell} \mathbb{E}[Y_{i,j}Y_{k,\ell}] = \sum_{i\ne j\ne k\ne \ell} \mathbb{E}[Y_{i,j}]\cdot\mathbb{E}[Y_{k,\ell}] = 0 \\
C &= \sum_{1\le i<j<k\le r} \mathbb{E}[Y_{i,j}Y_{i,k}] \le \sum_{1\le i<j<k\le r} \mathbb{E}[I_{i,j}I_{i,k}] = \sum_{1\le i<j<k\le r}\sum_{x=1}^{n} p_X(x)^3 \\
&= \binom{r}{3}\sum_{x=1}^{n} p_X(x)^3 \le \frac{\sqrt{3}}{2}\left(\binom{r}{2}\|p_X\|_2^2\right)^{3/2}
\end{aligned}
$$

Hence

$$
\begin{aligned}
\mathbb{E}\left[\left(\sum_{1\le i<j\le r} Y_{i,j}\right)^2\right] &= A + B + 6C \\
&\le \binom{r}{2}\cdot\|p_X\|_2^2 + 0 + 6\cdot\frac{\sqrt{3}}{2}\left(\binom{r}{2}\|p_X\|_2^2\right)^{3/2} \\
&\le 6\left(\binom{r}{2}\|p_X\|_2^2\right)^{3/2}
\end{aligned}
$$

Applying Chebyshef's inequality to $\hat{\theta}$ yields

$$
\begin{aligned}
P\left(\left|\hat{\theta} - \|p_X\|_2^2\right| \ge \epsilon\cdot\|p_X\|_2^2\right) &\le \frac{\mathrm{Var}[\hat{\theta}]}{\epsilon^2\cdot\|p_X\|_2^4} \\
&\le \frac{\frac{1}{\binom{r}{2}^2} 6\left(\binom{r}{2}\|p_X\|_2^2\right)^{3/2}}{\epsilon^2\cdot\|p_X\|_2^4} = \frac{6\|p_X\|_2^3}{\binom{r}{2}^{1/2}\epsilon^2\cdot\|p_X\|_2^4} \\
&\le \frac{6}{(r/2)\|p_X\|_2\epsilon^2} & \binom{r}{2}^{1/2} \le \frac{r}{2} \\
&= \frac{12}{r\|p_X\|_2\epsilon^2} \le \frac{12}{r(1/\sqrt{n})\epsilon^2} & \|p_X\|_2^2 \ge \frac{1}{n}
\end{aligned}
$$

For

$$\begin{aligned} P\left(\left|\hat{\theta} - \|p_X\|_2^2\right| \geq \epsilon \cdot \|p_X\|_2^2\right) &\leq \frac{1}{3} \\ \iff \frac{12}{r(1/\sqrt{n})\epsilon^2} &\leq \frac{1}{3} \\ \iff \frac{36\sqrt{n}}{\epsilon^2} &\leq r \end{aligned}$$

□

Uniform-Test(p_X, n) method description:

1. Run Estimate-$\|p_X\|_2^2$ with $r \geq 36\frac{\sqrt{n}}{\epsilon^2}$ to get estimate $\hat{\theta}$ s.t

$$P\left(\left|\hat{\theta} - \mathbb{E}\left[\hat{\theta}\right]\right| \geq \epsilon \cdot \|p_X\|_2^2\right) \leq \frac{1}{3}.$$

2. If $\hat{\theta} \geq \frac{1+\alpha}{n}$, then **Reject**.

3. Otherwise, **Accept**.

Theorem 4.2.2. (**Correctness of** Uniform-Test(p_X, n))

1. If $X \sim U(n)$, then Uniform-Test(p_X, n) returns **Accept** w/ $P($**Accept**$) \geq 2/3$.

2. If X is ϵ-far from $Y \sim U(n)$, then Uniform-Test(p_X, n) returns **Reject** w/ $P($**Reject**$) \geq 2/3$.

Proof. We have two cases:

1. **Case** $X \sim U(n)$. Assume that $X \sim U(n)$, then it follows that

$$\|p_X\|_2^2 = \sum_{x=1}^{n} p_X(x)^2 = \frac{1}{n}.$$

   By theorem ??, we have two cases:

- **Case** $\hat{\theta} \geq \|p_X\|_2^2$. So

$$\begin{aligned} & P\left(\left|\hat{\theta} - \mathbb{E}\left[\hat{\theta}\right]\right| \geq \epsilon \cdot \|p_X\|_2^2\right) \leq \frac{1}{3} \\ \iff & P\left(\hat{\theta} \geq (\epsilon + 1) \cdot \|p_X\|_2^2\right) \leq \frac{1}{3} \end{aligned}$$

  By the Complement law

$$P\left(\hat{\theta} < (\epsilon + 1) \cdot \|p_X\|_2^2\right) \geq \frac{2}{3}.$$

  Since $\|p_X\|_2^2 = 1/n$, it follows that $P(\textbf{Accept}) \geq 2/3$.

- **Case** $\hat{\theta} < \|p_X\|_2^2$. Note that

$$\hat{\theta} < \|p_X\|_2^2 = \frac{1}{n} < \frac{1+\alpha}{n}.$$

  Hence $P(\textbf{Accept}) = 1 \geq 2/3$.

2. **Case** X is ϵ-far from $Y \sim U(n)$. We proceed by proving the Contrapositive, that is $P(\textbf{Reject}) \leq 2/3 \implies X$ is ϵ-close (not ϵ-far) from $Y \sim U(n)$. Let us assume that $P(\textbf{Reject}) \leq 2/3$, that is

$$P\left(\hat{\theta} > \frac{1+\alpha}{n}\right) < \frac{2}{3}. \tag{*}$$

By theorem ?? and the Complement law we note that (*only if* $\hat{\theta} < \|p_X\|_2^2$)

$$P\left(\hat{\theta} > (1-\epsilon) \cdot \|p_X\|_2^2\right) \geq 2/3.$$

Since (*) is strictly decreasing (opposite to c.d.f), it follows that

$$(1-\epsilon) \cdot \|p_X\|_2^2 < \frac{1+\alpha}{n} \iff \|p_X\|_2^2 < \frac{1+\alpha}{n(1-\epsilon)}.$$

Recall that

$$\|p_X - p_Y\|_2^2 = \|p_X\|_2^2 - \frac{1}{n}.$$

Fuck this shit...

□

4.3 Online Algorithms

4.3.1 The Secretary Problem

Definition 4.3.1. (**Problem**) Suppose we have n candidates whose positions X_i on $(\Omega, \mathcal{F}, P)$. The positions form a uniform random permutation of $[n]$.

Select a candidate w/ the constraint that decision is made as soon as the candidate is seen and cannot be reverted.

Preliminary Approaches

- **Take First**. Select the first candidate.

  Let the discrete random variable $X_1, \ldots, X_n$ on $(\Omega, \mathcal{F}, P)$ be position of best candidate.

  $$P(\text{Selected best candidate}) = P(X_1 = 1) = \frac{1}{n}.$$

- **Explore-Exploit**. Explore the first $n/2$ candidates, then select the first person that is better than first $n/2$ (or if all best in first half, then select last person).

  Let the discrete random variables $X_1, \ldots, X_n$ on $(\Omega, \mathcal{F}, P)$ be positions of candidates.

  $$\begin{aligned} P(\text{Selected best candidate}) &= P(X_2 \leq n/2, X_1 > n/2) \\ &\quad + P(X_3 \leq n/2, X_1 > n/2, X_2 > X_1) + \cdots \\ &\geq P(X_2 \leq n/2, X_1 > n/2) \\ &= P(X_2 \leq n/2)P(X_1 > n/2 \mid X_2 \leq n/2) > \frac{1}{4} \end{aligned}$$

Optimal Strategy

- Explore the first $x - 1$ candidates, then select the first candidate $i \geq x$ which is better than first $i - 1$ candidates.

Let the discrete random variables $X_1, \ldots, X_n$ on $(\Omega, \mathcal{F}, P)$ be positions of candidates.

$$\begin{aligned}
P(\text{Selected best candidate}) &= \sum_{i=x}^{n} P(\text{Select } i \cap X_1 = i) \\
&= \sum_{i=x}^{n} P(\text{Select } i \mid X_1 = i) \cdot P(X_1 = i) \\
&= \frac{1}{n} \sum_{i=x}^{n} P(\text{Select } i \mid X_1 = i)
\end{aligned}$$

Note that we only select candidate i iff second best of the $i-1$ is in the $x-1$ candidates. Hence

$$\begin{aligned}
P(\text{Selected best candidate}) &= \frac{1}{n} \sum_{i=x}^{n} P(\text{2nd best of first } i-1 \leq x-1 \mid X_1 = i) \\
&= \frac{1}{n} \sum_{i=x}^{n} \frac{x-1}{i-1} \\
&= \frac{x-1}{n} \sum_{i=x-1}^{n-1} \frac{1}{i}
\end{aligned}$$

Note that

$$\int_{x-1}^{n} \frac{\mathrm{d}t}{t} \leq \sum_{i=x-1}^{n-1} \frac{1}{i} \leq \int_{x-2}^{n-1} \frac{\mathrm{d}t}{t}$$

$$\iff \frac{x-1}{n} \ln \frac{n}{x-1} \leq P(\text{Selected best candidate}) \leq \frac{x-1}{n} \ln \frac{n-1}{x-2}$$

Differentiating wrt $u = x-1$ the lower bound yields

$$\frac{\mathrm{d}}{\mathrm{d}u} \frac{u}{n} \ln \frac{n}{u} = \frac{1}{n} (\ln n - \ln u - 1)$$

Setting the derivative to zero, yields a maxima for the lower bound, when $\ln u = \ln n - 1 = \ln n/e$. Hence optimal when $x - 1 = n/e$.

4.3.2 The Secretary Problem With Payoff

- In this variant, every candidate has a value $X_i \sim U[0, 1]$.

- **Goal**: Maximize the expectation of the selected candidate.
- **Strategy**: Explore first $x-1$ candidates, then select the first candidate $i \geq x$ which is better than first $i-1$ candidates.

Let the continuous random variables $X_1, \ldots, X_n \sim U[0,1]$ model the "values" of the candidates (in order of sampling). The value of the best candidate Y_t out of the first t candidates is

$$Y_t = \max\{X_1, \ldots, X_t\}.$$

We note that

$$\begin{aligned}
F_{Y_t}(x) &= P(Y_t \leq x) \\
&= \prod_{i=1}^{t} P(X_i \leq x) \\
&= \prod_{i=1}^{t} \int_0^x \mathrm{d}u \\
&= x^t
\end{aligned}$$

Hence by the fundamental theorem of calculus, we have

$$f_{Y_t}(x) = tx^{t-1}.$$

So by the definition of expectation, we have

$$\mathbb{E}[Y_t] = \int_0^1 tx^t \, \mathrm{d}x = \left[\frac{t}{t+1}x^{t+1}\right]_0^1 = \frac{t}{t+1}$$

The expected value of the person chosen is

$$V_x(n) = \sum_{t=x}^{n-1} P\left(\text{Select candidate } t\right) \mathbb{E}[\text{Candidate } t] + P(\text{Select candidate } n)\mathbb{E}[\text{Candidate } n]$$

Note that

$$\begin{aligned} P(\text{Select candidate } t) &= P(\text{Select } t)P(\text{Don't select } x \text{ to } t-1) \\ &= \frac{1}{t} \cdot \prod_{s=x}^{t-1} \left(\frac{s-1}{s} \right) = \frac{1}{t} \frac{x-1}{t-1} \\ P(\text{Select candidate } n) &= P(\text{Don't select } x \text{ to } n-1) = \prod_{s=x}^{n-1} \left(\frac{s-1}{s} \right) = \frac{x-1}{n-1} \\ \mathbb{E}\left[Y_t\right] &= \frac{t}{t+1} \\ \mathbb{E}\left[\text{Candidate } n\right] &= \frac{1}{2} \end{aligned}$$

So

$$\begin{aligned} V_x(n) &= (x-1) \sum_{t=x}^{n-1} \frac{1}{(t-1)(t+1)} + \frac{1}{2} \frac{x-1}{n-1} \\ &= \frac{x-1}{2} \left[\frac{1}{n-1} + \sum_{t=x}^{n-1} \frac{1}{t-1} - \frac{1}{t+1} \right] \\ &= \frac{x-1}{2} \left[\frac{1}{n-1} + \frac{1}{x-1} + \frac{1}{x} - \frac{1}{n-1} - \frac{1}{n} \right] \\ &= \frac{1}{2} \left(2 - \frac{1}{x} - \frac{x-1}{n} \right) \end{aligned}$$

Taking the derivative of $V_x(n)$ wrt x yields

$$\frac{\mathrm{d}}{\mathrm{d}x} V_x(n) = \frac{1}{2} \left(\frac{1}{x^2} - \frac{1}{n} \right)$$

Setting the derivative to zero, yields a maxima when $x = \sqrt{n}$, since $x \in \mathbb{Z}^+ \implies x \in \{\lfloor \sqrt{n} \rfloor, \lceil \sqrt{n} \rceil\}$

4.3.3 The Odds Algorithms

- Let $I_1, \ldots, I_n$ be a sequence of independent random indicator variables on $(\Omega, \mathcal{F}, P)$.

- Let $r_1, \dots, r_n$ be the **odds** s.t for all $1 \leq i \leq n$
$$r_i = \frac{P(I_i = 1)}{P(I_i \neq 1)} = \frac{p_i}{1 - p_i}.$$
- What is the probability that after trial x, there is exactly one success?
$$P\left(\sum_{i=x}^{n} I_i = 1\right) = \sum_{i=x}^{n} P(I_i = 1) \prod_{i \neq j}^{n} P(I_j \neq 1) = \sum_{i=x}^{n} r_i \prod_{i=x}^{n} (1 - p_i)$$
- If $\sum_{i=x}^{n} r_i < 1$ then probability of success decreases. Hence find largest x s.t. $\sum_{i=x}^{n} r_i \geq 1$.
- **Optimal Strategy**: Ignore everything before xth trial, then **stop** at first success.

Example 4.3.1. (**Classical Secretary Problem**) Let us define for all $1 \leq i \leq n$ random indicator variable I_i on $(\Omega, \mathcal{F}, P)$ s.t
$$I_i = \begin{cases} 1 & \text{If candidate } i \text{ is best candidate in first } i - 1 \\ 0 & \text{Otherwise} \end{cases}.$$
We first wish to show that $I_1, \dots, I_n$ are independent, that is for all $\mathcal{I} \in \mathcal{P}\{I_1, \dots, I_n\}$, $\mathcal{I}$ is an independent set. Let $\mathcal{I} = \{I_{i_1}, \dots, I_{i_k}\} \in \mathcal{P}(\{I_1, \dots, I_n\})$ be arbitrary. We wish to show that for all $j_1, \dots, j_k \in \{0, 1\}$
$$P(I_{i_1} = j_1, \dots, I_{i_k} = j_k) = \prod_{\ell=1}^{k} P(I_{i_\ell} = j_\ell).$$
Let $j_1, \dots, j_k \in \{0, 1\}$ be arbitrary. Since each permutation of the first i_ℓ candidates is equally likely, independent of whether candidate i_ℓ is better than candidate $i_{\ell'}$. So we have
$$\begin{aligned} P(I_{i_1} = j_1, \dots, I_{i_k} = j_k) &= P(I_{i_1} = j_1 \mid I_{i_2} = j_2, \dots, I_{i_k} = j_k) \cdot P(I_{i_2} = j_2, \dots, I_{i_k} = j_k) \\ &= P(I_{i_1} = j_1) \cdot P(I_{i_2} = j_2, \dots, I_{i_k} = j_k) \\ &\vdots \\ &= \prod_{\ell=1}^{k} P(I_{i_\ell} = j_\ell) \end{aligned}$$

Hence $I_1, \ldots, I_n$ are independent. Hence $r_i = 1/(i-1)$. Now let us consider largest x s.t $\sum_{i=x}^{n} \frac{1}{i-1} \geq 1$. We have

$$
\begin{aligned}
\sum_{t=x}^{n} \frac{1}{t-1} &\geq \int_{x}^{n+1} \frac{\mathrm{d}t}{t-1} \\
&\geq [\ln t - 1]_{x}^{n+1} \\
&\geq \ln \frac{n}{x-1} = 1
\end{aligned}
$$

Hence $x - 1 = n/e$.