CS 70 Notes

Henry Yan

January 2023

Random

$$\sum_{i=1}^{n} \frac{1}{i} \approx \ln n + \gamma_E,$$

where $\gamma_E = 0.5772\ldots$ is Euler's constant.

Basic Notation and Propositional Logic

$\mathbb{N}$ includes 0.

$A \subset B$ means A is a **proper subset** of B.

$A \subseteq B$ means A is a subset of B, where A can equal B.

$A \cup B$ is the **union** of A and B.

$A \cap B$ is the **intersection** of A and B.

$B \backslash A = B - A$ is called the **relative complement**, or the **set difference**, of A in B. In simpler terms, this is set of elements in B but not in A.

The **Cartesian product**, or **cross product**, of two sets A, B is denoted by $A \times B = \{(a,b)|a \in A, b \in B\}$.

The **power set** of A is denoted by $P(A)$ is the set of all subsets of A.

Statements can be written like $(\forall a \in \mathbb{N})(a^2 \in \mathbb{N})$. This statement is true.

$P \wedge Q$ is **conjunction** and means "P and Q."

$P \vee Q$ is **disjunction** and means "P or Q."

$\neg P$ is the **negation** of P, or "not P."

The **law of the excluded middle** states that, for any proposition P, either P is true or $\neg P$ is true (but not both).

A **tautology** is a propositional form that is always true regardless of the truth values of the propositions used. Conversely, a **contradiction** is a propositional form that is always false.

$P \implies Q$ is **implication** and means if "P, then Q." Here, P is called the **hypothesis**, and Q is called the

conclusion.

IMPORTANT: An implication $P \implies Q$ is false iff when P is true, Q is false.

Logically, $P \implies Q \equiv \neg P \vee Q$, where $\equiv$ is used to denote logical equivalency.

The following are all equivalent to $P \implies Q$.

1. if P, then Q
2. Q if P;
3. P only if Q;
4. P is sufficient for Q;
5. is necessary for P;
6. unless not P.

Given an implication $P \implies Q$, we can also define its

1. **Contrapositive**: $\neg Q \implies \neg P$
2. **Converse**: $Q \implies P$

De Morgan's Laws:

1. $\neg(P \wedge Q) \equiv (\neg P \vee \neg Q)$
2. $\neg(P \vee Q) \equiv (\neg P \wedge \neg Q)$

Proof by contraposition applies to statements of the form $P \implies Q$, whereas proof by contradiction is used to prove a statement P and reach a clearly false conclusion.

Stable Matching

The Propose-and-Reject Algorithm: This algorithm works by jobs proposing to their most desirable candidate and then the candidates reject all but their favourite offer. The next day, each job proposes to it's next most desirable candidate. On the second last day, at least one candidate will have no offer, and on the last day, every candidate will have exactly one offer, of which they are unable to reject as it is their only (remaining) offer.

This algorithm seeks to find a **stable matching**, which is a matching in which no pair (a pair of a job and a candidate) of a job and candidate would rather work with each other over their current pairing. Such a pair is called a **rogue couple**.

However, this algorithm does not working when there is no asymmetry, for example when trying to pair up $2n$ roommates. In this case, no stable matching exists.

Well-Ordering Principle: If $S \subseteq \mathbb{N}$ and $S \neq \varnothing$, then S has a smallest element.

Theorem 4.2. The matching output by the Propose-and-Reject algorithm is job/employer optimal.

Optimal Job for a Candidate: For a given candidate C, the optimal job for C is the highest ranked job on C's preference list that C could be paired with in any stable matching.

The **pessimal** candidate for a job is the lowest ranked candidate that it is ever paired with in some stable matching.

Theorem 4.3: If a matching is employer/job optimal, then it is also candidate pessimal.

However, a stable matching could be both employer and job optimal.

Graph Theory

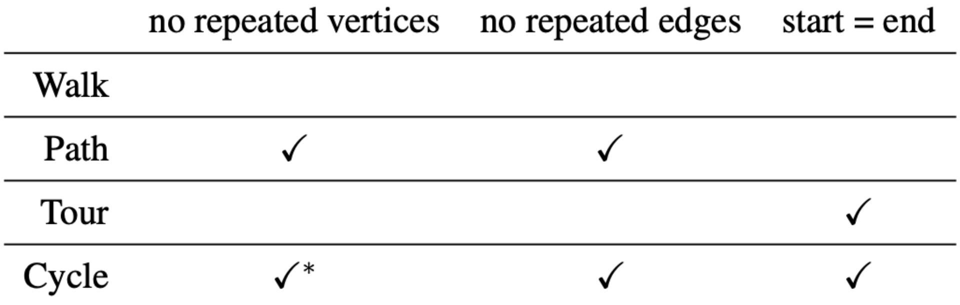
For a graph $G = (V, E)$, V is the set of vertices or nodes, and E is the set of edges. A graph G is said to be **simple** if it is undirected and does not have any loops or multiple edges.

We see that for an unordered edge $e = \{u, v\}$, e is **incident** on vertices u, v. We also say that u, v are **adjacent** or **neighbors**.

A **bridge** is a edge that has the same face on both sides of it. In other words, a bridge is a edge that will make another **connected component** if removed.

The number of edges incident to a vertex v is called the **degree** of v. A vertex will $deg(v) = 0$ is called an **isolated** vertex.

When considering a directed graph, we distinguish between **in-degree** and **out-degree**, where in-degree is the number of edges to u, and out-degree is the number of edges coming from u.



	no repeated vertices	no repeated edges	start = end
Walk			
Path	✓	✓	
Tour			✓
Cycle	✓*	✓	✓

Figure 1: Graph Theory Words

A graph is said to be **connected** if ∃ a path between any two distinct vertices.

An **even degree graph** is a graph in which all vertices have an even degree.

A **Hamiltonian tour** in G is a tour in G in which each vertex is used exactly once.

An **Eulerian walk** is a walk that uses every edge exactly once. If an Eulerian Walk is closed (start = end), it is then called an **Eulerian tour**.

Euler's Theorem: An undirected graph $G = (V, E)$ has an Eulerian tour iff G is even degree and connected (except possibly for isolated vertices).

The following are all definitions of a **tree** G:

1. G that is connected and acyclic (contains no cycles).
2. G is a connected graph with 1 more vertex than edge ($v = e + 1$).
3. Deleting any edge from G will cause the graph to become disconnected.
4. Adding any edge to G will create a cycle.

A **bipartite** graph is one where the vertices can be split into 2 groups and edges only go between the two groups. Bipartite graphs are 2-colorable, and vice versa. A graph that does not contain any odd length cycles is bipartite.

In a bipartite graph $G = (V, E)$, $|E| \leq 2|V| - 4$.

Euler's Formula: For every connected planar graph, $v + f = e + 2$.

$\sum_{i=1}^{f} s_i = 2e$, where s_i is the number of sides of face i.

$$3f \leq 2e, e \leq 3v - 6$$

Kuratowski's Theorem: A graph is non-planar iff it contains K_5 or $K_{3,3}$. K_5 is the complete (every node is connected to every other node) graph with 5 vertices, and $K_{3,3}$ is the complete bipartite graph separated into 2 groups of 3 vertices.

The **dual** of a graph G can be drawn by drawing a node on every face of G, then connecting all such nodes where their corresponding faces share an edge in G. Let the dual of G be G^*, then $(G^*)^* = G$. The dual of a tetrahedron is itself.

This is important as this allows us to see that "coloring a political map so that no two countries who share a border have the same color" is the same problem as "coloring the vertices of a planar graph (the dual of the political map) so that no two adjacent vertices have the same color."

Four-Colour Theorem: Any planar map can be colored using at most 4 colors.

Rooted Tree

A rooted tree is a tree with a designated **root** node at the top. The bottom-most nodes are called the **leaves** and any nodes between the root and leaves are called **internal nodes**. A root cannot be a leaf.

Leaves are any vertex of degree 1. The **depth** d of the tree is the length of the longest path from the root to a leaf. A good way to think about a rooted tree is to think about it as being grouped into layers or **levels**, where the k-th level for $k \in \{1, 2, \ldots, d\}$ is the set of vertices which are connected to the root by exactly k edges. The root is on level 0.

Hyper-Cubes

The vertex set of the n-dimensional hypercube $G = (V, E)$ is given by $\{0, 1\}^n$, where $\{0, 1\}^n$ represents the set of all n-bit strings. In other words, each vertex is labeled by a unique n-bit binary string. Two vertices x, y of the hypercube are connected by edge $\{x, y\}$ iff x and y differ in exactly one bit position.

The $(n+1)$ dimension hypercube can be formed by connecting all the corresponding vertices of two n dimension hypercubes (label one n dimension subcube $0x$ and the other $1x$).

The total number of edges in an n-dimensional hypercube is $n2^{n-1}$.

Modular Arithmetic

For $f : A \to B$:

1. **Onto** means surjective, every $b \in B$ has a pre-image $a \in A$.
2. **One-to-one** means injective, for $a, a' \in A$, if $f(a) = f(a')$, then $a = a'$.

A bijection is a function for which every $b \in B$ has a unique pre-image $a \in A$.

Lemma: For a finite set $A, f : A \to A$ is a bijection if there is an inverse function $g : A \to A$ such that for all $x \in A$, $g(f(x)) = x$.

The inverse of an element $a \in \mathbb{Z}_p$ is denoted by a^{-1}, not $\frac{1}{a}$.

Theorem 6.2: Let m, x be positive integers such that $\gcd(m, x) = 1$. Then x has a multiplicative inverse modulo m, and it is unique $\pmod{m}$.

Theorem 6.3: Let $x \geq y > 0$. Then $\gcd(x, y) = \gcd(y, x \pmod{y})$.

Euclid's Algorithm is just using Theorem 6.3 repeatedly until we reach 0 or 1 in one of the gcd values. $\gcd(x, 0) = x, \gcd(x, 1) = 1$.

The Fundamental Theorem of Arithmetic: The Fundamental Theorem of Arithmetic states that any positive integer greater than 1 can be expressed uniquely as a product of primes, up to a ordering of factors.

Claim: Let x, y, z be positive integers such that $\gcd(x, y) = 1$. If $x|yz$, then $x|z$.

Chinese Remainder Theorem: For m, n with $\gcd(m, n) = 1$ that there is exactly one $x \pmod{mn}$ that satisfies the equations:

$$x \equiv a \pmod{n} \text{ and } x \equiv b \pmod{m}.$$

The proof follows from the existence of inverses of n and m respectively modulo m and n, which holds when $\gcd(n, m) = 1$.

Theorem 7.2 (Fermat's Little Theorem): For a prime p and any $a \in \{1, 2, \ldots, p-1\}$,

$$a^{p-1} \equiv 1 \pmod{p}.$$

Theorem 7.3 (Prime Number Theorem): Let $\pi(n)$ denote the number of primes less than or equal to n, then for $n \geq 17$,

$$\pi(n) \geq \frac{n}{\ln n},$$

and as $n \to \infty$, $\pi(n) = \dfrac{n}{\ln n}$.

Public Key Cryptography

If Alice wants to send a message x to Bob, then she will encrypt the message using some encryption function E and send Bob the message $E(x)$. Bob will have a decryption function D, and applying D on $E(x)$ will give you $D(E(x)) = x$.

In this scenario, any third party, call her Eve, will not have access to the message x even if she intercepts the message $E(x)$, she will not be able to get x as she doesn't have the decryption function D. Everyone has access to a public key E. **Theorem 7.1**: Under the above definitions of the encryption and decryption functions E and D, we have $D(E(x)) \equiv x \pmod{N}$ for every possible message $x \in \{0, 1, ..., N-1\}$.

RSA

In RSA, pick two large primes p, q and define $N = pq$. Note that p, q are private. Let e be any number relatively prime to $(p-1)(q-1)$; typically e is a small value such as 3. Then Bob's public key is (N, e). Bob's private key is $d \equiv e^{-1} \pmod{(p-1)(q-1)}$.

Encryption and decryption:

1. Alice will send the message $E(x) \equiv x^e \pmod{N}$, where x is the message that Alice wants Bob to receive.
2. Bob will receive $y = E(x)$, and Bob will needs to compute $D(y) \equiv y^d \equiv x^{ed} \equiv x \pmod{N}$ to get the encrypted message.

Polynomials

Property 1: A non-zero polynomial of degree d has at most d roots.

Property 2: Given $d+1$ pairs $(x_1, y_1), \dots, (x_{d+1}, y_{d+1})$, with all x_i distinct, there is a unique polynomial $p(x) = a_d x^d + \cdots + a_1 x + a_0$ of degree (at most) d such that $p(x_i) = y_i$ for $1 \leq i \leq d+1$.

Property 2 tells us that 2 points uniquely determines a line, 3 points uniquely determine a quadratic, etc.

In the definition of property 2, the $d+1$ values $(a_d, \dots, a_1, a_0)$ are called the **coefficient representation** of p.

Polynomial Division: For polynomial p of degree d, we divide by $q(x)$ of degree $\leq d$ to get

$$p(x) = q'(x)q(x) + r(x),$$

where $q'(x)$ is the quotient and $r(x)$ is the remainder, where $\deg(r) < \deg(q)$.

A polynomial $p(x)$ of degree d with distinct roots $a_1, \dots, a_d$ can be written as $p(x) = c(x - a_1) \dots (x - a_d)$, where c is a real number.

Polynomial/Lagrange interpolation

Denote $\Delta_i(x) = \dfrac{\prod_{j \neq i}(x - x_i)}{\prod_{j \neq i}(x_i - x_j)}$.

Then

$$p(x) = \sum_{i=1}^{d+1} y_i \Delta_i(x).$$

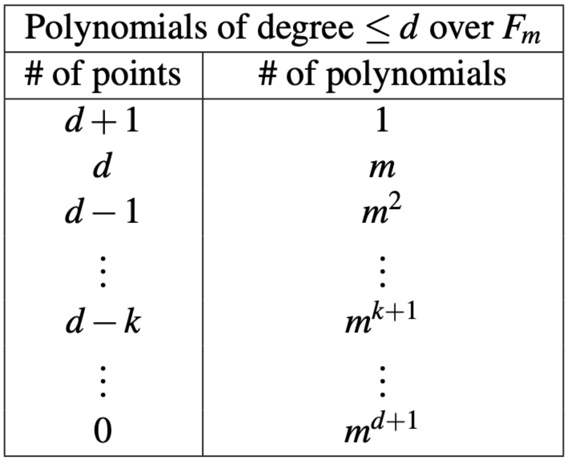
We can think of $(\Delta_i(x))$ as a natural basis for the space of all polynomials whose values are specified at the points $\{x_j\}$. We see that $\Delta_i(x)$ is determined solely by x_j, not y_j. We then find p by taking the coefficient in front of each basis in the sum to be the value of $y_i\Delta_i(x)$.

This process still works when working over the rationals or complex numbers, as these sets are also closed under addition, subtraction, multiplication, and division. As a result, natural numbers and integers don't work.

The $d+1$ values $(y_0, \ldots, y_{d+1})$ are called the **value representation** of p.

Finite Fields

A consequence of this is that polynomial interpolation works over a set of residues $\pmod p$ for some prime p, which we denote by F_p (field $\pmod p$)) or $GF(p)$ (for Galois Field).



Polynomials of degree $\leq d$ over F_m	
# of points	# of polynomials
$d+1$	1
d	m
$d-1$	m^2
$\vdots$	$\vdots$
$d-k$	m^{k+1}
$\vdots$	$\vdots$
0	m^{d+1}

Figure 2: Number of Polynomials of Degree d Given $\leq d+1$ Points

Secret Sharing

Pick a random polynomial $P(x)$ of degree $k-1$ such that $P(0) = s$ and give $P(1)$ to the first official, $P(2)$ to the second, ..., $P(n)$ to the n-th.

Then exactly k keys can find the specific polynomial $P(x)$ and determine s, whereas anything less than k keys convey 0 information about s.

Thus, over a finite field F_m, for some prime m, with insufficient information, there are $m = |\{0, 1, \ldots, m-1\}|$ possible values for s.

Error-Correcting Codes

A **codeword** is what is transmitted and has redundancy by construction while the **message** is something that the user gives us and does not necessarily contain any redundancy.

Distance (Optional - Note 9)

We define the **Hamming distance** of two strings (vectors) of length n, s and r, to be the number of positions in which the two strings differ.

The **minimum distance** of a code is defined as the distance between the two closest codewords. For distinct messages m, m', the minimum distance of the code is $\min\{d(c(m)), d(c(m'))\}$.

The greater the minimum distance, the more protection one can build-in against errors or erasures. When the minimum distance is $k+1$, one can protect against at most k errors. This is easy to intuitively see as a code with minimum distance k, meaning any two codewords have at most k differences, we cannot protect against k errors as if all the differences are error'd out, then we are left with the same word.

If $d-1$ positions are changed, we should, in principle, be able to decode the unique codeword that is less than $d/2$ from the received string.

Erasure Error

Erasure errors are errors where packets are lost during transmission.

Assume that the initial message has n packets and at most k are lost during transmission. Note that it is assumed that each of the packets have headers, or labels, so it is known which packets were lost during transmission.

For example, a 32-bit string message can be regarded as a number between 0 and $2^{32}-1$, then we can choose any prime $q > 2^{32}$, so we are working over the finite field $GP(q)$.

Denote the message by $m_1, \dots, m_n$, where $m_i \in GF(q)$ for the chosen prime q. Then

1. $\exists$ a unique polynomial p of degree $n-1$ such that $P(i) = m_i, i = 1, \dots, n$.
2. The additional k packets can be generated by evaluating p at points $n+j$. Then the transmitted codeword is $c_1 = p(1), \dots, c_{n+k} = p(n+k), n+k \leq q$. However, q is generally assumed to be very large for the sake of security, so the final inequality is essentially negligible.
3. We can reconstruct p from any n values of c_i via polynomial interpolation.

General Errors

Again, assume that Bob wants to send some message denoted by $m_1, \dots, m_n$ to Alice, and there are at most k general errors. Then we need $n+2k$ packets to ensure that the message is not corrupted.

Then, Bob's task is to find a polynomial P of degree $n-1$ such that $P(i) = r_i$ for at least $n+k$ values of i.

An **error-locator polynomial** is defined as $E(x) = (x-e_1)(x-e_2)\dots(x-e_k)$. Then $P(i)E(i) = r_iE(i)$, for $1 \leq i \leq n+2k$.

Let $Q(x) = P(x)E(x)$, which is degree $n+k-1$, then

$$Q(x) = a_{n+k-1}x^{n+k-1} + \cdots + a_1x + a_0$$
$$\text{and } E(x) = x^k + b_{k-1}x^{k-1} + \cdots + b_1x + b_0.$$

Then expanding $P(i)E(i) = r_iE(i)$ gives

$$a_{n+k-1}i^{n+k-1} + a_{n+k-2}i^{n+k-2} + \cdots + a_1i + a_0 = r_i(i_k + b_{k-1}i^{k-1} + \cdots + b_1i + b_0) \pmod{q}.$$

This is a set of $n+2k$ linear equations, one for each value of i. Then, we can compute $P(x) = \dfrac{Q(x)}{E(x)}$.

Counting

For $n \in \mathbb{N}$,

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}.$$

Properties of the binomial function:

1. $\sum_{k=0}^{n} (-1)^k \binom{n}{k} = 0$
2. $\binom{n}{k+1} = \binom{n-1}{k} + \binom{n-2}{k} + \cdots + \binom{k}{k}$
3. $2^n = \binom{n}{0} + \cdots + \binom{n}{n}$

Stars and Bars: When trying to sort k indistinguishable items into n distinguishable containers, there are $\binom{n+k-1}{k} = \binom{n+k-1}{n-1}$ possibilities, assuming each bucket is allowed to be empty.

If buckets are not allowed to be empty, then there are $\binom{k-1}{n-1}$ possibilities.

Principle of Inclusion-Exclusion: For a finite set A and subsets $A_1, A_2 \subset A$,

$$|A_1 \cup A_2| = |A_1| + |A_2| - |A_1 \cap A_2|.$$

Inclusion-Exclusion: For a finite set A and subsets $A_1, A_2 \subset A$,

$$|A_1 \cup \cdots \cup A_n| = \sum_{k=1}^{n} (-1)^{k-1} \sum_{S \subseteq \{1,\ldots,n\}:\ |S|=k} |\cap_{i \in S} A_i|$$

Derangements

A permutation with no fixed points is called a **derangement**.

For $n \geq 3$, the number of derangements D_n of $\{1, \ldots, n\}$ satisfies

$$D_n = (n-1)(D_{n-1} + D_{n-2})$$

Closed form for D_n

$$D_n = n! \sum_{k=0}^{n} \frac{(-1)^k}{k!}$$

Countability

We say a set S is **countable** if there exists a bijection between S and $\mathbb{N}$.

The Cantor-Bernstein Theorem: If $|A| \leq |B|$ and $|B| \leq |A|$, then there exists a one-to-one function $f : A \to B$ and one-to-one function $g : B \to A$, respectively. The existence of such f, g implies that there exists a bijection $h : A \to B$.

The power set of some set S, denoted $P(S)$, is the set of all subsets of S. More formally,

$$P(S) = \{T : T \subseteq S\}.$$

For some finite set S, $|P(S)| = 2^{|S|}$.

$|P(\mathbb{N})| > |\mathbb{N}|$

Cantors diagonalization argument

We claim that $\mathbb{R}[0, 1]$ is uncountable.

Proof. Write every element in $\mathbb{R}[0, 1]$ as a non-terminating decimal, i.e. $0.1 = 0.099\ldots$. Then assume, for contradiction, that there exists a bijection $f : \mathbb{N} \to \mathbb{R}[0, 1]$ such that

$$f(0) = 0.\mathbf{1}823650285\ldots$$

$$f(1) = 0.2\mathbf{8}4627102\ldots$$

Then for each $f(n), n \in \mathbb{N}$, construct a new number such that you take the ith digit in the decimal of $f(i)+2 \pmod{10}$. Then this new number is not in $(f(i)), i \in \mathbb{N}$, so f is not surjective, and is thus not a bijection. □

We say any infinite countable set S, for example the natural numbers, $|S| = \aleph_0$, or aleph null.

This cardinality, $|\mathbb{R}[0, 1]| = |P(\mathbb{N})|$, is known as c, the "cardinality of the continuum." So $2^{\aleph_0} = c > \aleph_0$.

Self-reference and Computability

"This statement is false" is a paradox. If the statement is true, then the statement is false, as stated in the statement itself. However, if the statement is false, then the statement is true, so we have a contradiction either way.

A program that prints itself is called a **quine**.

The recursion theorem states that given any program $P(x, y)$, we can always convert it to another program $Q(x)$ such that $Q(x) = P(x, Q)$, i.e., Q behaves exactly as P would if its second input is the description of the program Q. In this example, we can consider Q a self-aware version of P, since Q essentially has access to it's own description.

$$\text{TestHalt}(P, x) = \begin{cases} \text{“yes”, if program } P \text{ halts on input } x \\ \text{“no”, if program } P \text{ loops on input } x \end{cases}$$

Theorem - (The Halting Problem): There does not exist a computer program TestHalt with the behavior specified above on all inputs (P, x).

Consider the following code:

```
Turing($P$)
if TestHalt($P, P$) = "yes," then loop forever
else halt
```

The easy halting program: We show that this problem is also unsolvable, as its solvabilty is the same as that of the harder halting problem.

$$\text{TestEasyHalt}(P) = \begin{cases} \text{“yes”, if program } P \text{ halts on input } 0 \\ \text{“no”, if program } P \text{ loops on input } 0 \end{cases}$$

```
Halt$(P, x)$
construct a program $P'$ that, on input 0, returns $P(x)$
return TestEasyHalt$(P')$
```

Godel's Incompleteness Theorem: At least one of the following is false:

1. Is arithmetic **consistent**?
2. Is arithmetic **complete**?

If it is possible to prove a proposition P and $\neg P$, then we say that arithmetic is inconsistent. Otherwise, we say arithmetic is consistent.

“Is arithmetic complete?” asks weather every true statement can be proved.

Sketch of Gödel's Proof

Suppose we have a formal system F, which consists of a list of axioms and rules of inference, and assume F is sufficiently expressive that we can use it to express all of arithmetic. Now suppose we can write the following statement: $S(F) =$ “This statement is not provable in F.” Now consider the cases of $S(F)$ being either probable or not probable.

Relation to Halting Problem

Let $S_{P,x}$ denote the proposition that “P halts on input x.” Then the following program shows that by the uncomputability of the halting problem, Godel's Incompleteness Theorem is true:

```
Search$(P, x)$
for every proof $q$:
if $q$ is a proof of $S_{P,x}$ then output “yes”
if $q$ is a proof of $\neg S_{P,x}$ then output “no”
```

Probability - pt. 1

A **probability space** is a sample space Ω, together with a **probability** $\mathbb{P}[w]$, often denoted $Pr[w]$, such that

1. $0 \leq \mathbb{P}[w] \leq 1$, for all $w \in \Omega$,
2. $\sum_{w \in \Omega} \mathbb{P}[w] = 1$.

Formally, an **event** A is just a subset of the sample space Ω, i.e., $A \subseteq \Omega$. Then, $\mathbb{P}[A] = \sum_{w \in A} \mathbb{P}[w]$.

The use of a probability space for computing probability is basically just saying that the probability of something occurring is the number of ways for that to occur divided by the number of items in the probability space.

$\mathbb{P}[w|B] = \frac{\mathbb{P}[w]}{\mathbb{P}[B]}$

For events $A, B \in \Omega$, the **conditional probability** of A given B is

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}.$$

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[B|A]\mathbb{P}[A]}{\mathbb{P}[B|A]\mathbb{P}[A] + \mathbb{P}[B|\overline{A}](1 - \mathbb{P}[A])}$$

We say that an event A is **partitioned** into n events $A_1, \ldots, A_n$ if

1. (all of A is contained) $A = A_1 \cup A_2 \cdots \cup A_n$, and
2. (mutually exclusive) $A_i \cap A_j = \emptyset$, for all $i \neq j$.

The **Total Probability Rule** states that for any event B and partition $A_1, \ldots, A_n$ of A,

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

Baye's Rule states that

$$\mathbb{P}[A_i|B] = \frac{\mathbb{P}[B|A_i]\mathbb{P}[A_i]}{\mathbb{P}[B]}.$$

Two events A, B in the same probability space are said to be **independent** if

$$\mathbb{P}[A \cap B] = \mathbb{P}[A]\mathbb{P}[B].$$

A corollary of the above definition is that for independent events A, B, $\mathbb{P}[A|B] = \mathbb{P}[A]$.

Events $A_1, \ldots, A_n$ are said to be **mutually independent** if for every subset $I \subseteq \{1, \ldots, n\}$ with size $|I| \geq 2$, $\mathbb{P}[\bigcap_{i \in I} A_i] = \prod_{i \in I} \mathbb{P}[A_i]$.

Events $A_1, \ldots, A_n$ are said to be **mutually independent** if for all $B_i \in \{A_i, \overline{A_i}\}$, $i = 1, \ldots, n$,

$$\mathbb{P}[\bigcap_{i=1}^{n} B_i] = \prod_{i=1}^{n} \mathbb{P}[B_i].$$

Events $A_1, \ldots, A_n$ are said to be **mutually exclusive**, i.e., $A_i \cap A_j = \emptyset$ for all $i \neq j$ if for all $B_i \in \{A_i, \overline{A_i}\}$, $i = 1, \ldots, n$,

$$\mathbb{P}[\bigcup_{i=1}^{n} A_i] = \sum_{i=1}^{n} \mathbb{P}[A_i].$$

Union Bound: Let $A_1, \ldots, A_n$ be events in some probability space. Then, for all $n \in \mathbb{N}$,

$$\mathbb{P}[\bigcup_{i=1}^{n} A_i] \leq \sum_{i=1}^{n} \mathbb{P}[A_i].$$

Inclusion-Exclusion: Let $A_1, \ldots, A_n$ be events in some probability space, where $n \geq 2$. Then, we have

$$\mathbb{P}[\bigcup_{i=1}^{n} A_i] = \sum_{k=1}^{n} (-1)^{k-1} \sum_{S \subseteq \{1, \ldots, n\} : |S| = k} \mathbb{P}[\bigcap_{i \in S} A_i]$$

A value X that depends on the outcome of a probabilistic experiment is called a **random variable**. For example, X could be the number of heads when flipping 4 fair coins.

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

Until further notice, we will restrict our attention to random variables that are discrete, i.e., they take values in a range that is finite or countably infinite.

We use "$X = a$" to denote the set of events in a sample space $\{w \in \Omega : X(w) = a\}$, where a is any number in the range of random variable X.

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

Note that the collection of events $X = a$, for $a \in A$, satisfy two important properties:

1. Any two events $X = a_1$ and $X = a_2$ with $a_1 \neq a_2$ are disjoint.
2. The union of all these events is equal to the entire sample space Ω. The average value of a random variable X is called the **expectation** of X.

A simple yet very useful probability distribution is the **Bernoulli distribution** of a random variable which takes value in $\{0, 1\}$:

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

where $0 \leq p \leq 1$. We say that X is distributed as a **Bernoulli random variable** with parameter p, and write $X \sim \text{Bernoulli}(p)$.

The **binomial distribution** with the distribution

$$\mathbb{P}[X = i] = \binom{n}{i} p^i (1 - p)^{n-i}, i = 0, 1, \ldots, n.$$

A random variable X with this distribution is called a **binomial random variable**, which we denote by $X \sim \text{Bin}(n, p)$.

For $p = \frac{1}{2}$, the binomial distribution is a bell curve.

The **hypergeometric distribution** with the distribution

$$\mathbb{P}[Y = k] = \frac{\binom{B}{k}\binom{N-B}{n-k}}{\binom{N}{n}}.$$

A random variable X with this distribution is called a **hypergeometric random variable**, which we denote by $Y \sim \text{Hypergeometric}(N, B, n)$.

The **joint distribution** for two discrete random variables X and Y is the collection of values $\{((a, b), \mathbb{P}[X = a, Y = b]) : a \in A, b \in B\}$, where A is the set of all possible values taken by X and B is the set of all possible values taken by Y.

When given a joint distribution for X and Y, the distribution $\mathbb{P}[X = a]$ for X is called the **marginal distribution** for X, and can be found by "summing" over the values of Y. That is,

$$\mathbb{P}[X = a] = \sum_{b \in B} \mathbb{P}[X = a, Y = b].$$

Random variables X, Y on the same probability space are said to be **independent** if the events $X = a$ and $Y = b$ are independent for all values a, b. In math terms, this is equivalent to

$$\mathbb{P}[X = a, Y = b] = \mathbb{P}[X = a]\mathbb{P}[Y = b], \forall a, b.$$

Formally, let X be a random variable on a sample space Ω with probability distribution $\mathbb{P}_X$. Then for a function f on the range of X, $f(X)$ is the random variable Y on the same sample space Ω, where $Y(w) = f(X(w))$, for $w \in \Omega$.

In terms of sample spaces, the event $Y = y$ is equivalent to the event $X \in f^{-1}(y)$. $f^{-1}(y) = \{x|f(x) = y\}$. When f is one-to-one, it is a bit simpler, $X = x$ is the same event as $Y = f(x)$.

$$\mathbb{P}_Y[Y = y] = \sum_{x:f(x)=y} \mathbb{P}_X[X = x].$$

Law of the Unconscious Statistician (LOTUS):

$$\mathbb{E}[f(x)] = \sum_x f(x)P_X[X = x].$$

Expectation

The expectation of a discrete random variable X is defined as

$$\mathbb{E}[X] = \sum_{a\in\mathcal{A}} a \cdot \mathbb{P}[X = a],$$

where the sum is over all possible values taken by the random variable.

Linearity of Expectation: For random variables X, Y on the same probability space:

1. $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$
2. $\mathbb{E}[cX] = c\mathbb{E}[X]$, for any constant c.

Statistics

For a random variable X with expectation $\mathbb{E}[X] = \mu$, the **variance** of X is defined to be

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

The square root $\sigma(X) := \sqrt{\text{Var}(X)}$ is called the **standard deviation** of X.

The **variance** of a random variable X is

$$\text{Var}(X) = \sum_{w\in\Omega} Pr[X = w](w - \mathbb{E}[w])^2 = \sum_{w\in\Omega} w^2 Pr[X = w] - \mathbb{E}[w]^2.$$

Theorem 16.1: For a random variable X with expectation $\mathbb{E}[X] = \mu$, we have $\text{Var}(X) = \mathbb{E}[X^2] - \mu^2$.

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

Theorem 16.2: For independent random variables X, Y, we have $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.

Theorem 16.3: For independent random variables X, Y, we have $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$.

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

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

where $\mu_X = \mathbb{E}[X]$ and $\mu_Y = \mathbb{E}[Y]$.

Covariance is a measure of direction and strength of correlation between two random variables.

Covariance Facts:

1. If X, Y are independent, then $\text{Cov}(X, Y) = 0$. However, the converse is not true.
2. $\text{Cov}(X, X) = \text{Var}(X)$.
3. Covariance is bilinear; i.e., for any collection of random variables $\{X_1, \ldots, X_n\}$, $\{Y_1, \ldots, Y_m\}$ and fixed constants $\{a_1, \ldots, a_n\}, \{b_1, \ldots, b_m\}$,

$$\text{Cov}(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{m} b_j Y_j) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j \,\text{Cov}(X_i, Y_j)$$

For general random variables X and $Y, \text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\,\text{Cov}(X, Y)$.

Suppose X and Y are random variables with $\sigma(X) > 0$ and $\sigma(Y) > 0$. Then, the **correlation** of X and Y is defined as

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma(X)\sigma(Y)}.$$

Theorem 16.4: For any pair of random variables X and Y with $\sigma(X) > 0$ and $\sigma(Y) > 0$,

$$-1 \leq \text{Corr}(X, Y) \leq 1.$$

Geometric Distributions

A **geometric random variable** X for which

$$\mathbb{P}[X = i] = (1 - p)^{i-1}p,$$

for $i = 1, 2, 3, \ldots$, is said to have the geometric distribution with parameter p. This is abbreviated as $X \sim \text{Geometric}(p)$.

Theorem 19.1 - (Tail Sum Formula): Let X be a random variable that takes values in $\{0, 1, 2, \ldots\}$. Then

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

Theorem 19.2: For $X \sim \text{Geometric}(p)$, we have $\mathbb{E}[X] = \dfrac{1}{p}$.

Theorem 19.3: For $X \sim \text{Geometric}(p)$, we have $\text{Var}(X) = \dfrac{1 - p}{p^2}$.

The Memoryless Property of geometric random variables is

$$\mathbb{P}[X > n + m | X > m] = \mathbb{P}[X > n].$$

If a card collector gets 1 of the n possible collectable cards in each cereal box they buy, what is the expected number of cereal boxes needed to get at least one copy of each of the n cards. Answer: $n(\ln n + \gamma_E)$, where $\gamma_E = 0.5772\ldots$ is Euler's constant.

Poisson Distribution

A **Poisson random variable** X for which

$$\mathbb{P}[X = i] = \frac{\lambda^i}{i!} e^{-\lambda},$$

for $i = 0, 1, 2, \ldots,$ is said to have the **Poisson distribution** with parameter λ. This is abbreviated as $X \sim \text{Poisson}(\lambda)$.

Theorem 19.4: For a Poisson random variable $X \sim \text{Poisson}(\lambda)$, we have $\mathbb{E}[X] = \lambda$ and $\text{Var}(X) = \lambda$.

Theorem 19.5: Let $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\mu)$ be independent Poisson random variables. Then, $X + Y \sim \text{Poisson}(\lambda + \mu)$. This is easily generalized to more than 2 mutually-independent Poisson random variables.

Theorem 19.6: Let $X \sim \text{Binomial}(n, \frac{\lambda}{n})$ where $\lambda > 0$ is a fixed constant. Then for every $i = 0, 1, 2, \ldots,$

$$\mathbb{P}[X = i] \to \frac{\lambda^i}{i!} e^{-\lambda} \text{ as } n \to \infty.$$

That is, the probability distribution of X converges to the Poisson distribution with parameter λ.

Concentration Inequalities and the Laws of Large Numbers

Theorem 17.1 - (*Markov's Inequality*): For a non-negative random variable X (i.e., $X(w) \geq 0$ for all $w \in \Omega$) with finite mean,

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

for any positive constant c.

Theorem 17.2 - (*Generalized Markov's Inequality*): Let Y be an arbitrary random variable with finite mean. Then, for any positive constants c and r,

$$\mathbb{P}[|Y| \geq c] \leq \frac{\mathbb{E}[|Y|^r]}{c^r}.$$

Theorem 17.3 - (*Chebyshev's Inequality*): For a random variable X with finite expectation $\mathbb{E}[X] = \mu$,

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

for any positive constant c.

Corollary 17.1: For any random variable X with finite expectation $\mathbb{E}[X] = \mu$ and finite standard deviation $\sigma = \sqrt{\text{Var}(X)}$,

$$\mathbb{P}[|X - \mu| \geq k\sigma] \leq \frac{1}{k^2},$$

for any constant $k > 0$.

Cantelli's Inequality: For a random zero-mean variable X,

$$P[X \geq a] \leq \frac{\sigma^2}{a^2 + \sigma^2}$$

The **confidence level** $1-\delta$ and the **error** ε must

$$\mathbb{P}[|X-\mu| \geq \varepsilon] \leq \frac{\mathrm{Var}(X)}{\varepsilon^2} \leq \delta,$$

where a 95% confidence interval or confidence level corresponds to $\delta = 5\% = 0.05 = \frac{1}{20}$. The error is the radius of the confidence interval.

Theorem 17.4 - (*Law of Large Numbers*): Let $X_1, X_2, \ldots, X_n$ be a sequence of i.i.d. (independent and identically distributed) random variables with common finite expectation $\mathbb{E}[X_i] = \mu$ for all i. Then, their partial sums $S_n = X_1 + X_2 + \cdots + X_n$ satisfy

$$\mathbb{P}[|\frac{1}{n}S_n - \mu| < \varepsilon] \to 1 \text{ as } n \to \infty,$$

for every $\varepsilon > 0$, however small.

The best mean squared error of estimator for X is $\mathbb{E}[X]$ and this estimate results in a mean squared error of $\mathrm{Var}(X)$.

Joint Distributions

The **joint distribution** for two discrete random variables X and Y is the collection of values $\{((a,b), \mathbb{P}[X = a, Y = b]) : a \in A, b \in B\}$, where A is the set of all possible values taken by X and B is the set of all possible values taken by Y.

When given a joint distribution for X and Y, the distribution $\mathbb{P}[X = a]$ for X is called the **marginal distribution** for X, and can be found by "summing" over the values of Y. That is,

$$\mathbb{P}[X = a] = \sum_{b \in B} \mathbb{P}[X = a, Y = b].$$

The **conditional probability** of $X = x$ given $Y = y$ is

$$\mathbb{P}[X = x|Y = y] = \frac{\mathbb{P}[X = x, Y = y]}{\mathbb{P}[Y = y]}.$$

The **conditional expectation** of X given $Y = y$ is defined naturally as

$$\mathbb{E}[X|Y = y] = \sum_{x \in A} x \cdot \mathbb{P}[X = x|Y = y].$$

The Law of Iterated Expectations states

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]] = \sum_{y \in B} \mathbb{E}[X|Y = y]\mathbb{P}[Y = y].$$

Wald's identity states that for $Y = X_1 + \cdots + X_n$, where the X_i are identical and independently distributed,

$$E[Y] = E[X_1]E[N].$$

Linear Regression Line: The *line of best fit* or *linear least squares estimate* of Y given X (denoted LLSE$(Y|X)$) is given by,

$$\mathcal{L}(X) = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}(X - E[X]) + E[Y] = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}X - \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}E[X] + E[Y].$$

That is the **correlation coefficient** squared is exactly the fraction by which the mean squared error in the linear regression estimator is less than the error in estimating Y by using the estimate $E[Y]$. Thus, square of the correlation coefficient tells us how much the variance is explained by a linear estimator given X.

When working in a **continuous** real-valued space, we use intervals, measured by their length or other geometric (higher dimensions using multiple integrals) methods to measure probability.

A **probability density function** (p.d.f.) for a real-valued random variable X is a function $f : \mathbb{R} \to \mathbb{R}$ satisfying:

1. f is non-negative: $f(x) \geq 0$ for all $x \in \mathbb{R}$.
2. The total integral of f is equal to 1: $\int_{-\infty}^{\infty} f(x)dx = 1$.

Then the **distribution** of X is given by

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

However, remember that f is not defined, it's not really anything as $\mathbb{P}[X = a] = 0$, for all values of a in the domain.

For a continuous random variable X, one often starts the discussion with the **cumulative distribution function** (c.d.f.) which is the function $F(x) = P[X \leq x]$. It is closely related to the probability density function for $X, f(x)$, as

$$F(x) = \mathbb{P}[X \leq x] = \int_{-\infty}^{x} f(z)dz.$$

Then the derivative of the CDF is the is the PDF:

$$f(x) = \frac{dF(x)}{dx}.$$

The **expectation** of a continuous r.v. X with probability density function f is

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

The **variance** of a continuous r.v. X with probability density function f is

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

A **joint density function** for two random variable X, Y is a function $f : \mathbb{R}^2 \to \mathbb{R}$ satisfying:

1. f is non-negative: $f(x, y) \geq 0$ for all $x, y \in \mathbb{R}$.
2. The total integral of f is 1: $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)dxdy = 1$.

Then the **joint distribution** of X and Y is given by:

$$\mathbb{P}[a \leq X \leq b, c \leq Y \leq d] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)dxdy, \text{ for all } a \leq b \text{ and } c \leq d.$$

Two continuous r.v.'s X, Y are **independent** if the events $a \leq X \leq b$ and $c \leq Y \leq d$ are independent for all $a \leq b$ and $c \leq d$:

$$\mathbb{P}[a \leq X \leq b, c \leq Y \leq d] = \mathbb{P}[a \leq X \leq b] \cdot \mathbb{P}[c \leq Y \leq d].$$

Theorem 21.1. The joint density of independent r.v.'s X and Y is the product of the marginal densities:

$$f(x,y) = f_X(x)f_Y(y) \text{ for all } x, y \in \mathbb{R}.$$

For $\lambda > 0$, a continuous random variable X with p.d.f

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

is called an **exponential random variable** with parameter λ, and we write $X \sim \text{Exp}(\lambda)$.

Theorem 21.2. Let X be an exponential random variable with parameter $\lambda > 0$. Then

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

For an exponential random variable X,

$$\mathbb{P}[X > t] = \int_t^\infty \lambda e^{-\lambda x} dx = -e^{-\lambda x}\Big|_t^\infty = e^{-\lambda t}.$$

For any $\mu \in \mathbb{R}$ and $\sigma > 0$, a continuous random variable X with PDF

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

is called a **normal random variable** with parameters μ and σ^2, and we write $X \sim N(\mu, \sigma^2)$. In the special case $\mu = 0$ and $\sigma = 1, X$ is said to have the **standard normal distribution**.

Lemma 21.1. If $X \sim N(\mu, \sigma^2)$, then $Y = \frac{X-\mu}{\sigma} \sim N(0,1)$. Equivalently, if $Y \sim N(0,1)$, then $X = \sigma Y + \mu \sim N(\mu, \sigma^2)$.

Theorem 21.3. For $X \sim N(\mu, \sigma^2)$,

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

Theorem 21.4. Let $X \sim N(0,1)$ and $Y \sim N(0,1)$ be independent standard normal random variables, and suppose $a, b \in \mathbb{R}$ are constants. Then $Z = aX + bY \sim N(0, a^2 + b^2)$.

Corollary 21.1. Let $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim \mu_Y, \sigma_Y^2$ be independent normal random variables. Then for any constants $a, b \in \mathbb{R}$, the random variable $Z = aX + bY$ is also normally distributed with mean $\mu = a\mu_X + b\mu_Y$ and variance $\sigma^2 = a^2\sigma_X^2 + b^2\sigma_Y^2$.

Theorem 21.5 - (*Central Limit Theorem*): Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with common finite expectation $\mathbb{E}[X_i] = \mu$ and finite variance $\text{Var}\, X_i = \sigma^2$. Let $S_n = \sigma_{i=1}^n X_i$. Then the distribution of $\frac{S_n - n\mu}{\sigma\sqrt{n}}$ converges to $N(0,1)$ as $n \to \infty$. In other words, for any constant $c \in \mathbb{R}$,

$$\mathbb{P}\left[\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq c\right] \to \frac{1}{\sqrt{2\pi}} \int_\infty^c e^{-x^2/2} dx, \text{ as } n \to \infty.$$

Note that the Central Limit Theorem applies for any random variable $X = X_i, \forall i$.