Contents

1 undecided

Sets, Experiment, and Probability

1.1.1 Experiment

Definition 1.1
Experiment is a repeatable task with well defined outcomes.

Example 1.1.1
Examples of Experiment include:

(a) Toss a coin

(b) Draw ball from a bag containing balls

(c) Choose three students from this class
Note: It need to specify if choosing with or without replacement.

Definition 1.2
Sample Space is the set of all possible outcomes of an experiment, usually denoted by S

Example 1.1.2
Following the examples of Experiment above:

(a) Coin toss has $S = \{H, T\}$

(b) Ball draw has $S =$ set of all balls in the ba

(c) Choose three students has $S =$ all groups of three students in the class

Definition 1.3
Outcome is the collection of possible outcomes of an experiment. It is a subset of sample space S

Example 1.1.3
One of the outcome of coin toss is $\{H\} \subset S$, which means coin lands heads.
If we randomly choose three students in the United States, then $\{Sarah, Eric, Josh\}$ can be an outcomes which means the students being selected are Sarah, Eric, Josh.

1.1.2 Basic Set Theory

Definition 1.4
Let A, B be two sets.
We say A is a **subset** of B if and only if for all $x \in A, x \in B$, denoted as $A \subset B$.
We say A **equals** to B if and only if $A \subset B$ and $B \subset A$, which means they have exactly the same elements in it.

Definition 1.5
The **universal set** is the set of all objects of interest, denoted as U.

Example 1.1.4
In statistic, it is typically the sample space S of an experiment. If the experiment is tossing a coin, then $\{H, T\}$ is the U in this case.

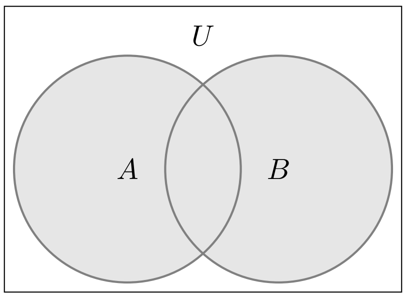
Definition 1.6
Let $\emptyset = \{x : x \neq x\}$. Then $\emptyset$ is a set with no elements in it, called **empty set** or **null set**

Note: Note that $\{\emptyset\}$ is not the empty set, rater it is a set containing one thing, that thing is the empty set.

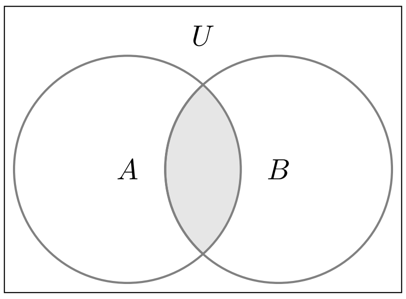
Definition 1.7
Let A, B be sets. Then
The **union** of A and B is the set $A \cup B = \{x : x \in A \text{ or } x \in B\}$
The **intersection** of A and B is the set $A \cap B = \{x : x \in A \text{ and } x \in B\}$
The set A minus set B is defined as $A \setminus B = \{x \in A : x \notin B\}$
The **complement** of A is the set $A^c = \{x : x \notin A\}$.

Venn Diagram is useful for visualizing all of the above and for a good intuitive understanding of when things are true and not true. In a Venn diagram we start with a rectangle that indicates the universal set and typically another two sets A and B.

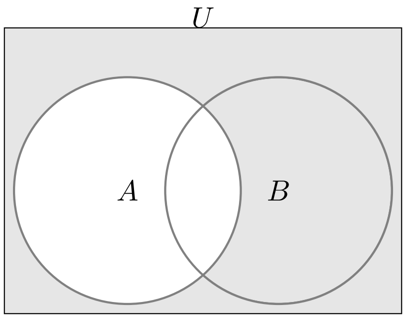
Example 1.1.5
Here is the shade of $A \cup B$



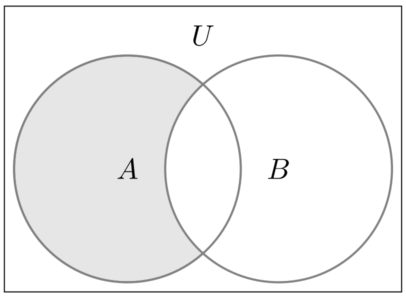
Example 1.1.6
Here is the shade of $A \cap B$



Example 1.1.7
Here is the shade of A^c



Example 1.1.8
Here is the shade of $A \setminus B$



Definition 1.8
Let A and B be sets. A and B are **disjoint** if and only if $A \cap B = \emptyset$ which means they are mutually exclusive.

Definition 1.9
Let $A_1, A_2, A_3, \cdots$ be sets. They are **pairwise disjoint** if $A_i \cap A_j = \emptyset$ for all $i \neq j$

Definition 1.10
Let $S, A_1, A_2, A_3, \cdots$ be sets. They are said to be a **partition** of S if

(a) $A_1, A_2, A_3, \cdots$ are pairwise disjoint

(b) $\bigcup_{i=1}^{\infty} A_i = S$

Definition 1.11
Let Γ be an indexing set that can be finite, countable infinite or uncountable infinite. Then, fix a set S, define $\{A_\alpha : \alpha \in \Gamma\}$ is a collection of subset of S indexed by Γ. Then, the **union** of $\{A_\alpha : \alpha \in \Gamma\}$ is define as

$$\bigcup_{\alpha \in \Gamma} A_\alpha = \{x \in S : x \in A_\alpha \text{ for some } \alpha \in \Gamma\}$$

The **intersection** of $\{A_\alpha : \alpha \in \Gamma\}$ is defined as

$$\bigcap_{\alpha\in\Gamma} A_\alpha = \{x \in S : x \in A_\alpha \text{ for all } \alpha \in \Gamma\}$$

1.1.3 Axioms of Probability

Definition 1.12
Let S be the sample space of an experiment. Given an outcome/event $A \subset S$, we are interested to calculate the chance or probability that A occurs, which denote as $P(A)$, or the **Probability of A**

Axiom 1.13
Axioms for a Probability function: A probability function P satisfies the following:

(i) $P(A) \geq 0$ for any event $A \subset S$

(ii) $P(S) = 1$ where S is the sample space.

(iii) For $A_1, A_2, A_3, \cdots$, the collection of pairwise disjoint events, we must have

$$\begin{aligned} P(A_1 \cup A_2 \cup A_3 \cup \cdots) &= P(A_1) + P(A_2) + P(A_3) + \cdots \\ &= \sum_{i=1}^{\infty} P(A_i) \end{aligned}$$

Theorem 1.14
A probability function P has the following properties:

(i) The probability of no outcome is zero, therefore $P(\emptyset) = 0$

(ii) Given an event $A \subset A$, A^c is set of all outcomes not in A. Then

  (a) $A^c \cap A = \emptyset$ and $A^c \cup A = S$
  (b) $P(A) + P(A^c) = P(S) = 1$
  (c) $O(A^c) = 1 - P(A)$

(iii) Given two events, $A, B \subset S$. Then

  (a) $A = (A \setminus B) \cup (A \cap B)$ Note that this is a disjoint union
  (b) $P(A) = P(A \setminus B) + P(A \cap B)$
  (c) $P(A\setminus) = P(A) - P(A \cap B)$

(iv) Given two events, $A, B \subset S$. Then

  (a) $A \cup B = (A \setminus B) \cup (A \cap B) \cap (B \setminus A)$. Note that this is pairwise disjoint union.
  (b)

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

Probability and Counting

Suppose

(a) S is the sample space of an experiment with finitely many outcomes, i.e S is a finite set.

(b) Every outcome in S is equally likely, i.e if $S = \{s1, s_2, \cdots, s_n\}$ then $P(\{s_i\}) = \frac{1}{n}$ for all $i = 1, 2, 3, \cdots, n$

Then, let an event $A \subset S$ be given. Since evey outcome in S is equally likely, then

$$P(A) = \text{Probability that A occurs} = \frac{n(A)}{n(S)}$$

where $n(A)$ = number of objects in A.
Note: In this Scenario, calculating probability is same as counting outcomes in an event, which is elementary but not easy!

Example 1.2.1
Here is concrete examples:

(a) Sample one elements from $\{A, B, C\}$, i.e $S = \{A, B, C\}$. Then $P(\{A\}) = P(\{B\}) = P(\{c\}) = \frac{1}{3}$, then every outcome is equally likely

(b) Sample one element from $\{A, A, B, C\}$, i.e $S = \{A, B, C\}$. Then $P(\{A\}) = \frac{2}{4}, P(\{B\}) = P(\{c\}) = \frac{1}{4}$, which means every outcome of this experiment is not equally likely.

1.2.1 Counting

Theorem 1.15
Fundamental Theorem of Counting
Suppose there are k tasks: $T_1, T_2, \cdots, T_k$ that can be performed in $n_1, n_2, \cdot, n_k$ ways respectively. Let T be the task of performing $T_1, T_2, \cdot, T_k$ sequentially. Then, the total number of ways to perform T is

$$\text{number of ways to perform T} = n_1 \times n_2 \times \cdots \times n_k$$

where n_i is number of ways to do T_i.
Typically we want to count the number of ways of selecting k objects from a set of n objects.

Example 1.2.2
In the set $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$, we might be interested in knowing the total number of ways one can choose 4 digits from this 10 digits, which has four possibilities.

	Without Replacement	With Replacement
Ordered	(1,2,4,5) different from (1,5,2,4) (1,1,2,5) is not possible	(1,2,4,5) different from (1,5,2,4) (1,1,2,5) is possible
Unordered	(1,2,3,4) is same as (4,3,2,1) (1,1,2,5) not possible	(1,2,3,4) is same as (4,3,2,1) (1,1,2,4) is possible

The **number of possible arrangement of size k from n objects 1. Without replacement and order matters** for each of the four possibilities is listed below

	Without Replacement	With Replacement
Ordered	${}^nP_k = \frac{n!}{n-k}$ is also called “n permute k”	n^k
Unordered	${}^nC_k = \frac{n!}{(n-k)!k!}$ is also called “n choose k”	${}^{n+k-1}C_k$

And we can use the Fundamental Theorem of Counting to prove!

(1) **Ordered Without Replacement**
Use the fundamental theorem of counting, we divide T, which is select k objects from a set of n distinct objects divide into

$$T : T_1 \to T_2 \to T_3 \to \cdots \to T_k$$

where T_i is select ith object. Then, we will got

$$n \times (n-1) \times (n-2) \times (n-3) \times \cdots \times (n-k+1)$$

Then, we got

$${}^nP_k = \frac{n!}{(n-k)!}$$

(2) **Unordered Without Replacement**
The first step is to calculate the ordered arrangements, which is

$${}^nP_k = \frac{n!}{(n-k)!}$$

Note that each ordered arrangement can be rearranged $k!$ times since we select k objects in total. Therefore, we need to get rid repeats by dividing by $k!$, which is

$$\frac{{}^nP_k}{k!} = \frac{n!}{(n-k)!k!}$$

So we got

$${}^nC_k = \binom{n}{k} := \frac{n!}{(n-k)!k!}$$

(3) **With Replacement**

The number of ways if choose k objects where order does matter and with replacement in n different things is simply

$$\begin{aligned} & n \times n \times n \times \cdots \times n \\ =& n^k \end{aligned}$$

(4) **With replacement and order does not matter**
We want number of unordered arrangements of size k from n objects with replacement. This can be reformulate as the number of ways to choose k "walls" from $(n+k-1)$ choices, i.e

$$^{n+k-1}C_k = \binom{n+k-1}{k} \text{ ways}$$

Example 1.2.3
We set $n = 10, k = 3$, where $S = \{0,1,2,3,4,5,6,7,8,9\}$.
Then we will have $n + k - 1 = 12$ walls

Then, for the event $(1,1,2)$, we will have (note that X means the wall that we choose)

1	X	X	2	X	3	4	5	6	7	8	9

For the event $(0,0,7)$, we will have

X	X	1	2	3	4	5	6	7	X	8	9

for the event $(5,9,7)$, we will have

1	2	3	4	5	X	6	7	X	8	9	X

Conditional Probability and Independence

1.3.1 Conditional Probability

Definition 1.16
Conditional Probability is the measure of the probability of an event A occurring, given that some other events B have already occurred, denoted by $P(A \mid B)$, the probability of A given B

Definition 1.17
Suppose $A, B \subset S, P(B) > 0$, Then

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

Example 1.3.1
Assume all outcomes in S are equally likely and that $|S| < \infty$, i.e S is finite. Suppose A, B are two non-empty subsets of S, and we want to calculate $P(A \mid B)$.

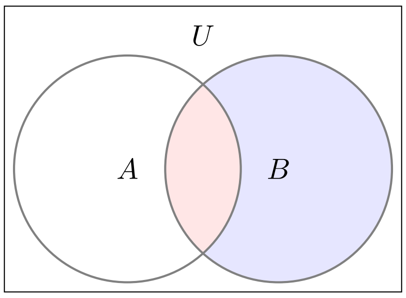
Note that $A \mid B$ means probability of A given B has occurred, the sample size we are looking at is reduceced to only those outcomes that are in B. Which means

$$P(A \mid B) = \frac{n(\text{outcomes in A given new sample space B})}{n\text{outcomes in new samplespace, i.e B}}$$

which can be rewritten as

$$\begin{aligned} P(A \mid B) &= \frac{n(A \cap B)}{n(B)} \\ &= \frac{n(A \cap B)}{n(B)} \frac{n(S)}{n(S)} \\ &= \frac{\frac{n(A \cap B)}{n(S)}}{\frac{n(B)}{n(S)}} \\ & \text{Since all outcomes are equally likely} \\ &= \frac{P(A \cap B)}{P(B)} \end{aligned}$$

Here is some visualization, the blue shaded area is the new sample space while the red shaded area is the outcomes in A assuming B has occurred, which is $A \cap B$



Theorem 1.18
Given $A, B \in S$ where $P(B) > 0$ and $P(A) < 0$. We want to calculate $P(A \cap B)$, the probability of events A and B occur. To do so, we can use the conditional probability, which states that

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

This implies that

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

Alternatively,

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

Note:

1. Conditioning is a very important tool
2. In the formula

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

The condition that $P(B) > 0$ and $P(A) > 0$ is important. Without it can lead to paradoxes.

1.3.2 Independence

Example 1.3.2
Before talking about the notion of independence, let's consider the following events

$$A = \text{Jonathan is carrying an umbrella}$$
$$B = \text{It is raining outside}$$
$$C = \text{It is sunny outside}$$

Then we can note that

$$P(A) < P(A \mid B) = P(\text{Jonathan carrying umbrella} \mid \text{raining outside})$$
$$P(A) > P(A \mid C) = P(\text{Jonathan carrying umbrella} \mid \text{sunny outside})$$

This show that knowing additional information can affect the probability of an outcome!

Now, Let D = had pasta for breakfast. We would expect the event D have no impact on Jonathan carrying an umbrella, i.e we can expect

$$P(A \mid D) = P(A)$$

We can also say that A and D are independent, A and B are dependent.

Definition 1.19
$A, B \subset S$ are **independent** if and only if

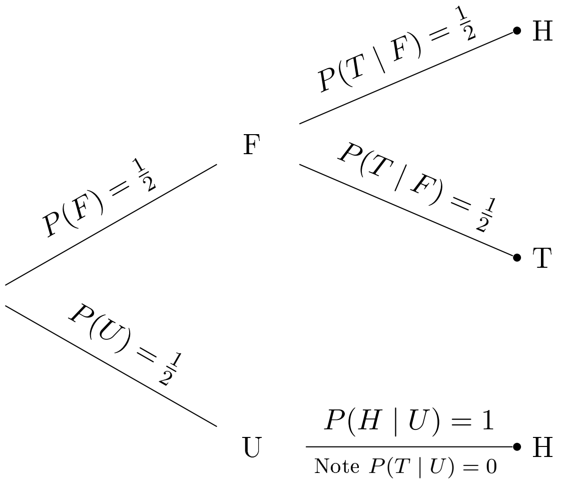
$$P(A \mid B) = P(A) \iff P(B \mid A) = P(B)$$

Alternatively, A, B are independent if and only if

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

Note: To show that two events are independent, we need to show $P(A \cap B) = P(A) \cdot P(B)$ or be able to calculate $P(A \mid B)$ or $P(B \mid A)$, which is not easy.

Example 1.3.3
Suppose there is a bag contains two coins, one is a fair coin denoted as F, and the other is a two headed coin denoted as U.
We randomly choose a coin from the bag and toss it once, which can be represented by the following tree diagram.



And we get the sample space $S = \{FH, FT, UH\}$, which is $F \cap H$, $F \cap T, U \cap H$ respectively.

Now

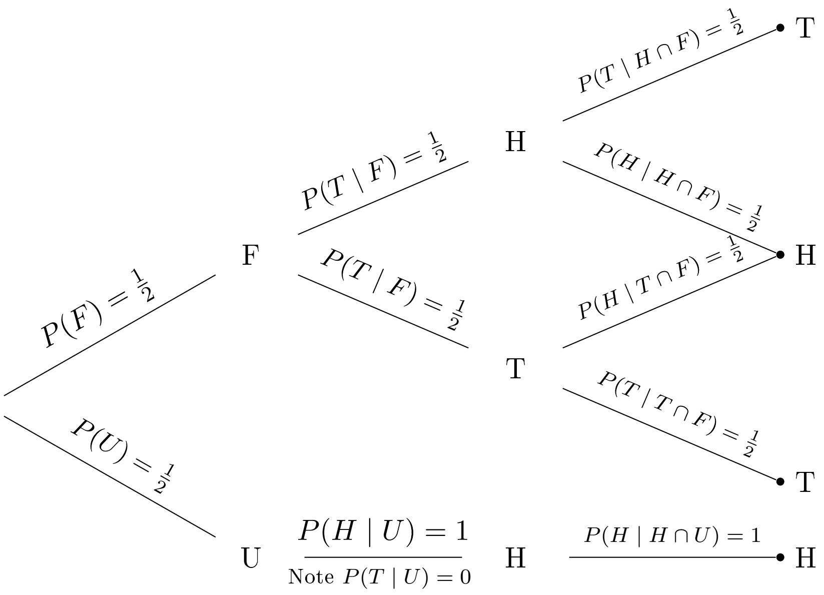
$$\begin{aligned} P(H) &= P(F \cap H) + P(U \cap H) \\ &= P(H \mid F) \cdot P(F) + P(H \mid U) \cdot P(U) \\ &= \frac{1}{2} \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = \frac{1}{4} + \frac{1}{2} = \frac{3}{4} \end{aligned}$$

Now, $P(H \mid F) = \frac{1}{2} \neq \frac{3}{4} = P(H)$, which implies that H and F are not independent events.
Also,

$$\begin{aligned} P(F|H) &= \frac{P(F \cap H)}{P(H)} \\ &= \frac{P(H \mid F) \cdot P(F)}{P(H)} \\ &= \frac{\frac{1}{2} \cdot \frac{1}{2}}{\frac{3}{4}} = \frac{1}{4} \cdot \frac{4}{3} = \frac{1}{3} \end{aligned}$$

Therefore, $P(F) = \frac{1}{2} > \frac{1}{3} = P(F|H)$, which also makes intutive sense that our confidence that the coin is fair should be reduce given the information that it has landed heads.

Then, suppose we randomly choose a coin from the bag and toss it independently two times, which can be visualized in tree diagram as the following:



Now we have $S = \{FHT, FHH, FTH, FTT, UHH\}$. Note that

$$\begin{aligned}&P(HHF) = P(H \cap (H \cap F))\\ =&P(H \mid H \cap F) \cdot P(H \cap F)\\ =&P(H \mid H \cap F) \cdot P(H \mid F) \cdot P(F)\\ =&\text{Product along path to HHF}\end{aligned}$$

Then, we want to calculate $P(HH)$, the probability of getting two heads

$$\begin{aligned}&P(HH)\\ =&P(FHH) + P(UHH)\\ =&P(H \mid HF) \cdot P(HF) + P(H \mid HU) \cdot P(HU)\\ =&\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot 1 \cdot 1 = \frac{1}{8} + \frac{1}{2} = \frac{5}{8}\end{aligned}$$

Also,

$$P(HH \mid F) = \frac{P(HHF)}{P(F)} = \frac{\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2}}{\frac{1}{2}} = \frac{1}{4}$$

Therefore, $P(HH) = \frac{5}{8} > \frac{1}{4} = P(HH \mid F)$, which mean we are less confident of getting two consecutive heads if we knew that the tossed coin is fair.
Also,

$$\begin{aligned}&P(F \mid HH)\\ &= \frac{P(FHH)}{P(HH)} = \frac{P(HH \mid F) \cdot P(F)}{P(HH)}\\ &= \frac{\frac{1}{4} \cdot \frac{1}{2}}{\frac{5}{8}} = \frac{1}{8} \cdot \frac{8}{5} = \frac{1}{5}\end{aligned}$$

Therefore, $P(F \mid HH) = \frac{1}{5} < \frac{1}{3} = P(F \mid H) < \frac{1}{2} = P(F)$, which implies that the confidence that the coin is fair decreases as observed consecutive heads increases from none to one and to two.

Bayes' Theorem

Theorem 1.20
Law of Total Probability
Let $\{A_1, A_2, A_3, \cdots, A_n\}$ be a partition for the sample space S, i.e

(a) $A_1, \cdots, A_n$ are mutually exclusive, $A_i \cap A_j = \emptyset \forall i \neq j$

(b) $A_1, \cdots, A_n$ are exhaustive, $\bigcup_{i=1}^{n} A_i = S$

Given any event $B \subseteq S$, we have

$$\begin{aligned}B =&B \cap S = B \cap \left(\bigcup_{i=1}^{n} A_i\right)\\ =&\bigcup_{i=1}^{n}(B \cap A_i)\end{aligned}$$

Then,

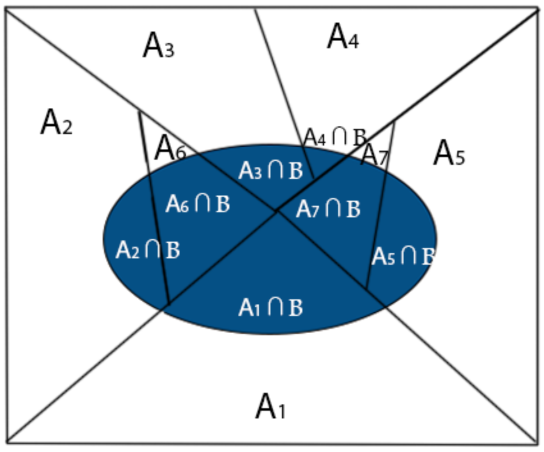
$$P(B) = P\left(\bigcup_{i=1}^{n}(B \cap A_i)\right) = \sum_{i=1}^{n} P(B \cap A_i)$$

Since $P(B \cap A_i) = P(B \mid A_i) \cdot P(A_i)$, then we can also rewrite $P(B)$ as the following

$$P(B) = \sum_{i=1}^{n} P(B \mid A_i) \cdot P(A)i)$$

This is the **law of total probability.**

Example 1.4.1
Given a sample space S and a set of $\{A_1, \cdots, A_7\}$ that is a partition of S and event $B \subseteq S$. Then, here is a graph representation of the law of total probability



Definition 1.21
Given $\{A_1, A_2, A_3, \cdots, A_n\}$ be a partition for the sample space S and any event $B \subseteq S$. Then, $P(A_i)$ and $P(B \mid A_i)$ for $i = 1, 2, 3, \cdots, n$ are called **prior probabilities**

Theorem 1.22
Given the prior probabilities, we can calculate the **posterior probabilities**, the $P(A_j \mid B)$, $j = 1, 2, 3, \cdots, j = 1, 2, 3, \cdots, n$.

$$\begin{aligned} P(A_j \mid B) &= \frac{P(A_j \cap B)}{P(B)} \\ &= \frac{P(B \mid A_j) \cdot P(A_j)}{\sum_{i=1}^{n} P(B \mid A_j) \cdot P(A_j)} \end{aligned}$$

Example 1.4.2
Suppose a city has two types of cabs, blue and green one. We define

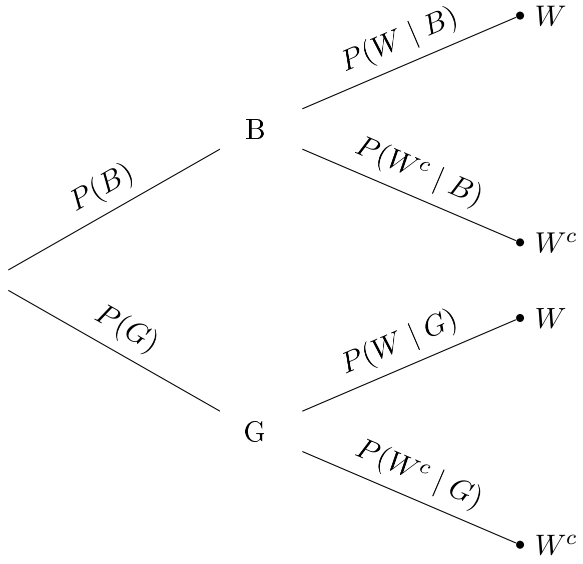
$$\begin{aligned} P(B) &\rightarrow \text{the proportion of blue cabs in the city} \\ P(G) = 1 - P(B) &\rightarrow \text{the proportion of green cabs in the city} \end{aligned}$$

Suppose there was a hit and run incident involving a cab in the city. With no additional information available, the probability of involved cab is the proportion of blue cabs in the city, i.e $P(B)$

Now, suppose we have a witness, We define

$$\begin{aligned} W &\rightarrow \text{event that witness says that the involved cab was a blue cab} \\ W^c &\rightarrow \text{event that witness says that the involved cab was a green cab} \end{aligned}$$

Then, we have the following tree diagram



Note that witness is correct only in the first and last case.

Then, we want to find the probability that the involved cab was a blue cab given that the witness says it was a blue cab, i.e the posterior probability $P(B \mid W)$. Using the Bayes' Theorem, we got

$$P(B \mid W) = \frac{P(W \mid B) \cdot P(B)}{P(W \mid B) \cdot P(B) + P(W \mid G) \cdot P(G)}$$

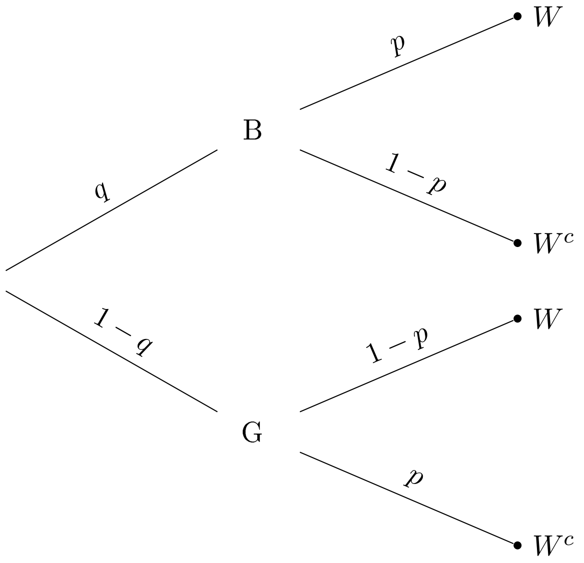
We now make prior choices for witness reliability. Let $p \in (0, 1)$ be the probability that witness makes a correct observation. Then, the events corresponding to correct observations by witness $W \mid B$, and $W^c \mid B$ has the following probabilities

$$P(W \mid B) = P(W^c \mid G) = p$$

Similarly, the events corresponding to incorrect observations by witness $W \mid G$, and $W^c \mid B$ has the following probabilities

$$P(W \mid G) = P(W^c \mid B) = 1 - p$$

Let $P(B) = q \in (0, 1)$. Now, we have the following update tree diagram



Then,

$$\begin{aligned} P(B \mid W) &= \frac{P(W \mid B) \cdot P(B)}{P(W \mid B) \cdot P(B) + P(W \mid G) \cdot P(G)} \\ &= \frac{p \cdot q}{p \cdot q + (1-p)(1-q)} \end{aligned}$$

This is a beautiful equation that leads to three important cases depending on prior assumption on witness reliability, i.e the value of p.

Case I) $P(B \mid W) = 0$, meaning witness testimony exonerates the blue cab.

$$\begin{aligned} P(B \mid 0) = 0 &\implies \frac{pq}{pq + (1-p)(1-q)} = 0 \\ &\implies pq = 0 \implies p = 0 \text{ or } q = 0 \end{aligned}$$

$p = 0$ means witness reliability is zero.
$q = 0$ means there are no blue cabs in the city.

Case II) $P(B \mid W) = 1$, meaning witness testimony results in blue cab involve with complete confidence.

$$\begin{aligned} P(B \mid W) = 1 &\implies \frac{pq}{pq + (1-p)(1-q)} = 1 \\ &\implies pq = pq + (1-p)(1-q) \implies (1-p)(1-q) = 0 \\ &\implies 1 - p = 0 \text{ or } 1 - q = 0 \\ &\implies p = 1 \text{ or } q = 1 \end{aligned}$$

$p = 1$ means witness is completely reliable.
$q = 1$ means there are only blue cabs in the city.

Case III) $P(B \mid W) = P(B)$, meaning witness testimony does not change our perception of the blue cab being involved.

$$\begin{aligned}
P(B \mid W) = P(B) &\implies \frac{pq}{pq + (1-p)(1-q)} = q \\
\therefore q \neq 0 &\implies p = pq + (1-p)(1-q) = pq + 1 - p - q + pq \\
&\implies 2p - 2pq = 1 - q \implies 2p(1-q) = (1-q) \\
\therefore (1-q) \neq 0 &\implies 2p = 1 \implies p = \frac{1}{2}
\end{aligned}$$

Therefore, witness testimony is independent of blue car being at faul corresponds to setting the prior witness reliability to the probability of a fair coin toss, i.e $p = \frac{1}{2}$

2 | undecided

Random Variables

Suppose we have an experiment with sample space S. We might be interested in a particular property of outcomes $\omega \in S$, as oppsed to being interested in the outcome ω itself.

Example 2.1.1
Let the experiment be tossing a two sided coin 5 times. Then

$$\begin{aligned} |S| &= 2s = 2^5 = 32 \\ S &= \{HHHHH, HHHHT, \cdots, TTTTH, TTTTT\} \end{aligned}$$

Then, given an $\omega \in S$, we might be interested in

(a) Number of heads in the outcome ω

(b) Are there at least two heads in ω?

(c) Is the number of heads equal to number of tails in ω>

(d) Are there more heads than tails in ω?

Example 2.1.2
Let the experiment be rolling two 6-sided dice. Then

$$\begin{aligned} S =& \{(1,1), \cdots, (1,6), \cdots, (6,1) \cdots (6,6)\} \\ |S| =& 36 \end{aligned}$$

For any $\omega \in S$, we might be interested in

(a) Is the sum greater than 5?

(b) Is there at least one even number?

(c) Is the sum divisible by 3?

(d) Is the max greater than 4?

All those questions in the example above are random variable.

Definition 2.1
Let S be the sample space of an experiment. A **random variable** X is a function whose domain is S and range is a new sample space (typically a subset of $\mathbb{R}$). i.e

$$\begin{aligned} X : S &\longrightarrow \mathbb{R} \\ \omega &\mapsto X(\omega) \end{aligned}$$

The image of X is the values of X on the outcomes in S, which specify a new sample space for the original experiment in the context of a certain property, i.e X extracts specific inromation about an outcome in S Suppose $c \in X(S)$, then

$$\begin{aligned} \{X = c\} &= \{\omega \in S : X(\omega) = c\} \\ &= \text{ All outcomes } \omega \in S \text{ which are mapped to c under X} \\ &= X^{-1}(c) \end{aligned}$$

Theorem 2.2
Let $\mathcal{X} \subseteq \mathbb{R}$ be the set of values of X, then

$$S = \bigcup_{c \in \mathcal{X}} \{X = c\}$$

The subsets $\{X = c\} \subseteq S$ form a partition of S

Example 2.1.3
Suppose we toss a coin 4 times independently. Then

$$S = \left\{\begin{array}{ccccc} HHHH & HHHT & HHTT & TTTH & TTTT \\ & HHTH & HTHT & TTHT & \\ & HTHH & HTTH & THTT & \\ & THHH & THTH & HTTT & \\ & & THHT & & \\ & & TTHH & & \end{array}\right\}$$

Let $X(\omega) =$ number of heads in the outcome ω. This is a random variable $X : S \to \mathbb{R}$. Then, we have

$$\begin{aligned} X(HHHT) &= 3 \\ X(HHHH) &= 4 \\ X(HTHT) = X(THTH) &= 2 \end{aligned}$$

Then let $\mathcal{X} =$ values of $X = \{0, 1, 2, 3, 4\}$. We can calculate $\{X = c\}$ for $c \in \mathcal{X}$.

$$\begin{aligned} \{X = 0\} &= \{TTTT\} \\ \{X = 1\} &= \{HTTT, THTT, TTHT, TTTH\} \\ \{X = 2\} &= \{HHTT, HTHT, HTTH, THTH, THHT, TTHH\} \\ \{X = 3\} &= \{HHHT, HHTH, HTHH, THHH\} \\ \{X = 4\} &= \{HHHH\} \end{aligned}$$

Then we can define the following sets with $\{X = c\}$ for $c \in \mathcal{X}$

$$\begin{aligned} E_1 = \{X \leq 2\} &= \text{ set of all outcomes with at most 2 heads} \\ &= \{X = 0\} \cup \{X = 1\} \cup \{X = 2\} \\ \\ E_2 = \{X > 1\} &= \text{ set of all outcomes with more than 1 heads} \\ &= \{X = 2\} \cup \{X = 3\} \cup \{X = 4\} \\ &= (\{X = 0\} \cup \{X = 1\})^c \\ \\ \{1 \leq X \leq 3\} &= \text{ set of outcomes with number of heads between 1 and 3} \\ &= \{X = 1\} \cup \{X = 2\} \cup \{X = 3\} \end{aligned}$$

We can observe that in this case

$$\bigcup_{c \in \mathcal{X}} \{X = c\} = S$$

We can also define probabilities of events $\{X = c\}$ as the following

$$P(\{X = c\}) = \sum_{\omega \in \{X = c\}} P(\omega)$$

For example,

$$\begin{aligned} P(X = 1) &= P(\text{ Getting exactly one head in 4 independent tosses}) \\ &= P(HTTT) + P(THTT) + P(TTHT) + P(TTTH) \end{aligned}$$

If we assume all outcomes in S are equally likely, we will have

$$\begin{aligned} P(HTTT) = P(THTT) = P(TTHT) = P(TTTH) = \frac{1}{16} \\ \therefore P(X = 1) = 4 \cdot \frac{1}{16} \end{aligned}$$

Similarly, we can calculate

$$\begin{aligned} P(X = 0) = P(X = 4) &= \frac{1}{16} \\ P(X = 2) &= \frac{6}{16} \\ P(X = 3) &= \frac{4}{16} \end{aligned}$$

These information can be organized into a table

$x \in \mathcal{X}$	0	1	2	3	4
$P(X = x)$	$\frac{1}{16}$	$\frac{4}{16}$	$\frac{6}{16}$	$\frac{4}{16}$	$\frac{1}{16}$

This is also called **probability distribution table for X**.
Note that

$$\begin{aligned} & P(X=0)+P(X=1)+P(X=2)+P(X=3)+P(X=4) \\ =& P\left(\bigcup_{c\in\mathcal{X}}\{X=c\}\right) \\ =& \frac{1}{16}+\frac{4}{16}+\frac{6}{16}+\frac{4}{16}+\frac{1}{16} \\ =& \frac{16}{16}=1 \\ =& P(S) \end{aligned}$$

We can also define a new random variable $Y(\omega)$ = proportion of heads in the outcome. Then,

$$\begin{aligned} Y =& \frac{\text{number of heads in an outcome}}{4} \\ =& \frac{X}{4} \end{aligned}$$

For example,

$$Y(HHTT) = T(TTHH) = \frac{2}{4}$$

Example 2.1.4
Let experiment be rolling a 6 sided dice 4 times. Then $|S| = 6^4 = 16 \times 81$.
Define $X(\omega)$ = sum of the numbers facing up. $\mathcal{X} = \{4, 5, \cdots, 24\}$. Then

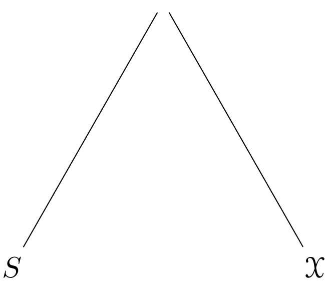
$$\begin{aligned} X(1,2,4,4) &= 11 \\ X(1,1,1,6) &= 9 \end{aligned}$$

Define $Y(\omega)$ = average of numbers in the sample space. Then, $\mathcal{Y} = \frac{n}{4} : n \in \mathbb{N}, 4 \leq n \leq 24$. This is because the minimum of the average is 1 and the maximum is 6. If one of the number increases by 1, the average will increases by $\frac{1}{4}$

Probability Distribution of Random Variables

2.2.1 Probability Function of Random Variables

Given an experiment with sample space S. For simplicity, assume $|S|$ is finite, i.e $S = \{S_1, S_2, \cdots, S_n\}$. Let $X : S \to \mathbb{R}$ be a random variable, let $\mathcal{X}$ = values of x = $\{x_1, x_2, \cdots, x_k\}$. Then there are two sample space for the experiment



Some simple outcomes of S are $\{S_j\}, j = 1, 2, 3, \cdots, n$ and note that Events $E \subseteq S$ are described in terms of $\{S_i\}$. Some simple outcomes of $\mathcal{X}$ are $\{X = x_i\} = \{s \in S : X(s) = x_i\} \subseteq S$. Events $E \subset S$ are described in terms of $\{X = x_i\}$
Then, let P be a **probability function** for S. We can use P to induce a probability function P_X on $\mathcal{X}$

Step 1)

$$P_X(\{x_i\}) := P(X = x_i) = \sum_{\omega\in\{X=x_i\}} P(\omega)$$

Note that $\{x_i\}$ are the outcomes in $\mathcal{X}$, $X = x_i$ means when the events ω in S has value of x_i.

Step 2) Suppose $E \subseteq \mathcal{X} = \{x_1, x_2, x_3, \cdots, x_k\}$, then $E = \{X = x_1\} \cup \cdots \cup \{X = x_r\}$

$$P_X(E) := P\left(\bigcup_{i=1}^{r}\{X = x_i\}\right) = \sum_{i=1}^{r} P(X = x_i)$$

For example, let $E = \{x_1, x_2, x_3\}$, then $E = \{X = x_1\} \cup \{X = x_2\} \cup \{X = x_3\} = \bigcup_{i=1}^{3} \{X = x_i\}$. Then $P_X(E) = \sum_{i=1}^{3} P(X = x_i)$

Note:

1. The random variable will denoted by upper case letters, $X, Y, \cdots, etc$
2. The values of the random variable will be denoted by lower cas letters, $x, y, \cdots, etc$
3. When there is no ambiguity, we will use the notation $P(X = x_i)$ instead of $P_X(X = x_i)$

2.2.2 Distribution of Random Variables

Definition 2.3

Given an experiment with sample space S. Let X be a random variable, P_X be a probability function for X. Then, the **cumulative distribution function** of X is

$$F_X(x) := P(X \leq x) \ \forall x \in \mathcal{X}$$
$$F_X : \mathbb{R} \longrightarrow \mathbb{R}$$
$$x \mapsto F_X(x) = P_X(X \leq x)$$

Note that when sample space is finite

$$F_X(x) = P(X \leq x) = \sum_{\substack{x_i \in \mathcal{X} \\ x_i \leq x}} P(X = x_i)$$

Example 2.2.1

Suppose we toss a fair coin 4 times. Let $X =$ number of heads in 4 tosses. We can calculate the following:

$$P(X = 0) = \frac{1}{16},\ P(X = 1) = \frac{4}{16},\ P(X = 2) = \frac{6}{16},\ P(X = 3) = \frac{4}{16},\ P(X = 4) = \frac{1}{16}$$

Using the above information,we can calculate $F_X(x)$ as

$$F_X(x) = \begin{cases} 0 & \text{if } \infty < x < 0 \\ \frac{1}{16} & \text{if } 0 \leq x < 1 \\ \frac{5}{16} & \text{if } 1 \leq x < 2 \\ \frac{11}{16} & \text{if } 2 \leq x < 3 \\ \frac{15}{16} & \text{if } 3 \leq x < 4 \\ \frac{16}{16} = 1 & \text{if } 4 \leq x < \infty \end{cases}$$

Note:

1. $F_X(x)$ has jumps at $x = 0, 1, 2, 3, 4$ and the jump equals $P(X = x_i)$
2. $F_X(x)$ is a non-decreasing function
3. $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$
4. $F(x) \geq 0 \ \forall x \in \mathbb{R}$

Theorem 2.4

A function $F(x)$ is a cumulative density function if and only if the following three conditions are satisifies:

(a) $\lim_{x \to -\infty} F(x) = 0$, $\lim_{x \to \infty} F(x) = 1$

(b) $F(x)$ is a non decreasing function

(c) $F(x)$ is right continuous. i.e $\lim_{x \to x_o^+} F(x) = F(x_o), \forall x_o \in \mathbb{R}$

Definition 2.5

Let X be a random variable. Then, X is **discrete** if $F_X(x)$ is a step function. Then, the probability mass function is

$$p_X(x) = P(X = x) \quad \forall x$$

If $x \notin \mathcal{X}, P(X = x) = 0$

Definition 2.6

Let X be a random variable. Then, X is **continuous** if $F_X(x)$ is a continuous function. Then, the probability density function $f_X(x)$ satisfies

$$\int_{-\infty}^{x} f_X(t)dt \quad \forall x$$

Theorem 2.7

$f_x(x)$ is a probability density function (or probability mass function) of a random variable if and only if

1. $f(x) \geq 0$ for all x
2. $\sum f_X(x) = 1$ for discrete random variable or $\int_{-\infty}^{\infty} f_X(x)dx = 1$ for continuous random variable

We can use cumulative density function and probability density function of a random variable X to calculate $P(a \leq X \leq b)$, P(X takes values in the interval $[a, b]$)

3 | Discrete Random Variable

Recall that X is said to be a discrete random variables if and only if F_X is a step function, which means $\mathcal{X}$ will be a discrete subset of $\mathbb{R}$, i.e $\mathcal{X}$ is a finite set, which means it is in bijection with $\{1, 2, 3, \cdots, N\}$ for some $N \in \mathbb{N}$, or countably infinite set, which means it is in bijection with $\mathbb{N} = \{1, 2, 3, 4, \cdots\}$

Uniform Discrete Random Variable

Definition 3.1
Fix $N \in \mathbb{N}$. The random variable X has **uniform discrete distribution** with parameter N if

$$\begin{aligned} \mathcal{X} &= \{1, 2, 3, \cdots, N\} \\ p_X(x) = P(X = x) &= \frac{1}{N} && \text{for } x = 1, 2, 3, \cdots, N \\ p_X(x) &= 0 && \text{for } x \notin \mathcal{X} \end{aligned}$$

Note:

1. $\sum_{x \in \mathcal{X}} P_X(x) = \underbrace{\frac{1}{N} + \frac{1}{N} + \cdots + \frac{1}{N}}_{\text{N times}} = N \cdot \frac{1}{N} = 1$, and $p_X(x) = \frac{1}{N} > 0 \ \forall x \in \mathcal{X}$
   Therefore p_X is a legitimate probability mass function

2. The distribution table for X is

x	1	2	3	4	$\cdots$	$N-1$	Na
$P(X = x)$	$\frac{1}{N}$	$\frac{1}{N}$	$\frac{1}{N}$	$\frac{1}{N}$		$\frac{1}{N}$	$\frac{1}{N}$

3. To calculate **E(X)**

$$\begin{aligned} E(X) &= \sum_{x \in X} x \cdot p_X(x) = \sum_{i=1}^{N} i \cdot \frac{1}{N} \\ &= \frac{1}{N} \sum_{i=1}^{N} i = \frac{1}{N} \cdot \frac{N(N+1)}{2} \\ &= \frac{N+1}{2} \end{aligned}$$

4. To calculate **V(X)** we need to know that $V(X) = E(X^2) - E(X)^2$, so we want to first calculate the $E(X^2)$

$$\begin{aligned} E(X^2) &= \sum_{x \in X} x^2 \cdot p_X(x) = \sum_{i=1}^{N} i^2 \cdot \frac{1}{N} \\ &= \frac{1}{N} \sum_{i=1}^{N} i^2 = \frac{1}{N} \cdot \frac{N(N+1)(2N+1)}{6} \\ &= \frac{(N+1)(2N+1)}{6} \end{aligned}$$

   Therefore, $V(X)$ is

$$\begin{aligned} V(X) = E(X^2) - E(X)^2 &= \frac{(N+1)(2N+1)}{6} - \left(\frac{N+1}{2}\right)^2 \\ &= \frac{(N+1)(N-1)}{12} = \left(\frac{N^2-1}{12}\right) \end{aligned}$$

5. At $\{a_1, a_2, \cdots, a_n\} \subseteq \mathbb{R}$, we define

$$\begin{aligned} h : \mathcal{X} &\longrightarrow \{a_1, a_2, \cdots, a_N\} \\ i &\mapsto a_i \end{aligned}$$

   Then if X has uniform distribution, $E(h(x)) = \frac{a_1 + a_2 + \cdots + a_N}{N}$, and $V(h(x)) =$ variance of the numbers $\{a_1, a_2, \cdots, a_N\} = \frac{1}{N} \sum_{i=1}^{N} (a_i - M)^2$

Binomial Distribution

Fix $N, M, n \in \mathbb{N}$. Suppose a bag has N balls that are identical except for color and M of them are read. Let the experiment be choosing n balls from the bag and define a random variable X = number of red balls in the sample of n balls.

If we want to calculate $p_X(x) = P(X = x)$, we first need to know if there is replacement or not. If it is sampling with replacement, then it is a binomial distribution with parameters: $n, p = \frac{M}{N}$. If without replacement, it is a hypergeometric distribution, which we will discuss later.

Definition 3.2

X is said to have **Binomial Distribution** with parameters: n, p if

$$\mathcal{X} = \{0, 1, 2, \cdots, n\}$$
$$p_X(x) = \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x}$$

Note:

1. $$\begin{aligned}\sum_{x\in\mathcal{X}} p_X(x) &= \sum_{x=0}^{n} \binom{n}{x} p^x \cdot (1-p)^{n-x} \\ &= (p + (1-p))^n = 1^n = 1\end{aligned}$$

   Therefore, p_X is a probability mass function.

2. To calculate **E(X)**

   $$E(X) = n \cdot p$$

3. To calculate **V(X)**

   $$V(X) = n \cdot p \cdot (1-p)$$

Hypergeometric Distribution

Reacall that if the choosing balls experiment mentioned in the binomial distribution is sampling without replacement, then it is a hypergeometric distribution with parameters: N, M, n.

Definition 3.3

X is said to have **hypergeometric distribution** with parameters: N, M, n if

$$\mathcal{X} = \{0, 1, 2, 3, \cdots, n\}$$
$$p_X(x) = P(X = x) = \frac{\binom{M}{x} \cdot \binom{N-M}{n-x}}{\binom{N}{n}}$$

Note that $P(X = x)$ is the probability of getting exactly x successes in a sample of size n without replacement from a population of size N with M successes.

Note:

1. x must satisfy the following condition

   $$max(0, n - N + M) \leq x \leq min(n, M)$$

   where n is the sample size, N is the population, M is the number of successes.

2. To calculate $E(X)$

   $$E(X) = n \cdot \frac{M}{N}$$

3. To calculate $V(X)$

   $$V(X) = \left(\frac{N-n}{N-1}\right) \cdot n \cdot \left(\frac{M}{N}\right) \cdot \left(1 - \frac{M}{N}\right)$$

4. If N is large enough, then $\frac{N}{M} \to p$, so we can use the Hypergeometric distribution(N, M, n) to approximate Binomial $(N, p = \frac{M}{N})$. Then if we set $p = \frac{M}{N}$

   $$\begin{aligned}&E(X) = n \cdot p \\ &V(X) = \frac{N-n}{N-1} \cdot n \cdot p \cdot (1-p) = n \cdot p \cdot (1-p) \text{ since } n \to \infty, \frac{N-n}{N-1} \approx 1\end{aligned}$$

Bernoulli Distribution

If an experiment has exactly two outcomes, and the probability of success is p and failure is $(1-p)$, then we said X is a bernolli random variable.

Definition 3.4
X is a **Bernolli** random variable with parameter p if

$$\mathcal{X} = \{0,1\}$$
$$p_X(\mathcal{X}) = \begin{cases} p & \text{if } x = 1 \\ (1-p) & \text{if } x = 0 \end{cases}$$

Then we have $E(X) = p$ and $V(X) = p \cdot (1-p)$

Poisson Distribution

Now, we want to study discrete random variables which have $\mathcal{X}$ to be infinte.

Definition 3.5
X is said to have **Poisson Distribution** with parameter $\lambda > 0$ if

$$\mathcal{X} = \{0,1,2,3,\cdots\}$$
$$p_X(x) = \frac{e^{-\lambda} \cdot \lambda^x}{x!} \text{ for } x \in \mathcal{X}$$

Note:

1. $$\begin{aligned} e^{\lambda} &= 1 + \lambda + \frac{\lambda^2}{2!} + \cdots \\ &= \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} \\ \implies 1 &= e^{-\lambda} \cdot \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = \sum_{x=0}^{\infty} e^{-\lambda} \cdot \frac{\lambda^x}{x!} = \sum_{x=0}^{\infty} p_X(x) \end{aligned}$$

   Also, $p_X(x) = \frac{e^{-\lambda} \cdot \lambda^x}{x!} \geq 0$ for all $x \in \mathcal{X}$. Therefore, p_X is a legitimate probability mass function.

2. The mean and variance are

   $$E(X) = M_x = \lambda$$
   $$V(x) = \sigma_X^2 = \lambda$$
   $$\sigma_X = \sqrt{\lambda}$$

3. The Poisson distribution is often used to model phenomena where we are waiting for events to happen.

Negative Binomial Distribution

Definition 3.6
X is said to have **Negative Binomial Distribution** with parameters r, p if

$$\mathcal{X} = \{0,1,2,3,\cdots\}$$
$$p_X(x) = \binom{x+r-1}{r-1} p^r \cdot (1-p)^x \text{ for } x \in \mathcal{X}$$

Note:

1. The experiment for Negative Binomial Distribution is performing Bernoulli trials with P(success) = p repeatedly until we get exactly r successes.
   Let X = number of failure that precede the r^{th} successes, then

   $$P(X = x) = P(\text{Exactly } x \text{ failures before } r^{th} \text{ successes})$$

And the total ways we can have exactly x failures before r^{th} successes is choosing x spots in $x+r-1$ possible places, since the last spot has to be a successes. Then, there are

$$\binom{x+r-1}{x} = \binom{x+r-1}{r-1} \text{ ways}$$

Therefore,

$$p_X(x) = \binom{x+r-1}{r-1} p^r \cdot (1-p)^x$$

2. The mean and variance are

$$E(X) = \mu_x = \frac{r(1-p)}{p}$$

$$V(x) = \sigma_X^2 = \frac{r(1-p)}{p^2}$$

3. We called it negative binomial distribution because

$$\binom{x+r-1}{x} = (-1)^x \binom{-r}{x}$$

Therefore,

$$p_X(x) = (-1)^x \binom{-r}{x} p^r (1-p)^x$$

Note that $\binom{-r}{x} p^r (1-p)^x$ is very similar to the probability mass function of Binomial Distribution

Geometric Distribution

Definition 3.7
We say X has **Geometric Distribution** with parameter p if

$$\mathcal{X} = \{0, 1, 2, 3, \cdots\}$$
$$p_X(x) = p(1-p)^{x-1} \text{ for } x \in \mathcal{X}$$

Note:

1. The Geometric distribution is a special case of Negative Binomial distribution when $r = 1$. It is the simplest of the waiting time distribution where X can be interpreted as the trial at which first successes occurs, which is "waiting for a success."

2. The mean and variances are

$$E(X) = \mu_x = \frac{1}{p}$$

$$V(x) = \sigma_X^2 = \frac{(1-p)}{p^2}$$

3. Sometimes it is used to moedl lifetime or time until failure of components

Relationships Between Discrete Distributions

3.8.1 Binomial and Hypergeometric

Let $N, M, n \in \mathbb{N}$ be given and set $p = \frac{M}{N}$. Then,

(a) bin$(x; n, p)$ is the probability mas function for binomial distribution with parameters:

$$n = \text{ sample size}$$
$$p = \text{ probability of success}$$

(b) hyper$(x; N, M, n)$ is the probability mass function for hypergeometric distribution with parameters

$$n = \text{ sample size}$$
$$M = \text{number of successes}$$
$$N = \text{population size}$$

Then, if $N, M \to \infty$, and $\frac{M}{N} \to p$, then hyper$(x; N, M, n)$ →bin$(x; n, p)$, i.e if sample size N is small compared to population size N, we can assume that samples are approximately independent.

3.8.2 Binomial and Poisson

Suppose

(a) $\mathrm{bin}(x; n, p)$ is the probability mas function for binomial Distribution with parameters:

$$\begin{aligned} n = & \text{ sample size} \\ p = & \text{ probability of success} \end{aligned}$$

(b) $\mathrm{pois}(x; \lambda)$ is the probability mass function for Poisson Distribution with parameters $\lambda > -$

Then, if $n \to \infty$ and $p \to 0$ such that $n \cdot p \to \lambda$, then then $\mathrm{bin}(x; N, M, n) \to \mathrm{pois}(x; n, p)$, i.e If n is large and p is small, $\mathrm{bin}(x; N, M, n) \approx \mathrm{pois}(x; n, p)$, or the Poisson Distribution is approximately Binomial fore rare events.

4 Continuous Random Variable

Introduction

Recall that X is a continuous random variable if and only if the cumulative density function F_x is a continuous function.
Let $f_X(x)$ be the probability density function of X which satisfies the following

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

Note:

1. $f_X(x)$ must satisfy:
   (a) $f_X(x) \geq 0$ for all $x \in (-\infty, \infty)$
   (b) $\int_{-\infty}^{\infty} f_X(x)dx = 1$
2. The expected value of X is

$$E(X) = \mu_x = \int_{-\infty}^{\infty} x f_X(x)dx$$

3. The variance of X is

$$V(X) = \sigma_X^2 = \int_{-\infty}^{\infty} (x - \mu_x)^2 f_X(x)dx$$

4. The standard deviation of X is

$$\sigma_X = \sqrt{\sigma_X^2}$$

5. Variance can be also written as following

$$V(X) = E(X^2) - (E(X))^2$$

6. There is one important identities associates with the expected value function

$$E(h(X)) = \int_{-\infty}^{\infty} h(x) f_X(x)dx$$

   for example,

$$E(aX + b) = aE(X) + b,\ V(aX + b) = a^2 V(X)$$

To calculate the probabilities for the events we want, we need to use the cumulative density function that is defined as the following

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

i.e $F_X(x)$ is the area under graph of f_X upto x.

Note:

1.

$$\begin{aligned} P(a \leq X \leq b) &= \int_a^b f_X(x)dx \ a \leq b \\ &= \int_{-\infty}^b f_X(x)dx - \int_{-\infty}^a f_X(x)dx \\ &= F_X(b) - F_X(a) \end{aligned}$$

2. The probability that a continious random variable take a particular value is zero

$$P(X = c) = P(c \leq x \leq c) = F_X(c) - F_X(c) = 0$$

3.

$$\begin{aligned} P(X > c) =& 1 - P(x \leq c) = 1 - F_X(c) \\ P(X \geq c) =& P(X > c) + P(X = c) = P(X > c) \end{aligned}$$

And we can use the cumulative density function F_X to calculate $f_X(x)$. If the derivative $F'_X(x)$ exist at x, then

$$F'_X(x) = f_X(x)$$

Let $p \in (0,1)$. Then the "$(100 \cdot p)th$" Percentile for the distribution of X is $\eta(p)$ and satisfies

$$p = F_X(\eta(p)) = \int_{-\infty}^{\eta(p)} f_X(x)dx$$

Let $\alpha \in (0,1)$. Then the the αth critical value for the distribution of X is $\mathcal{X}_\alpha$ and satisfies

$$\alpha = P(X > X_\alpha) = 1 - F_X(x\alpha)$$

Note that $(100p)th$ percentile is same as $(1-p)6h$ critical value.

Uniform Continuous Distribution

Let $[A, B] \subseteq \mathbb{R}$ be a closed interval where $A < B$. Then we can define the probability density function as the following

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

Then, the cumulative density function of the uniform continuous distribution is

$$F(x) = \begin{cases} 0 & \text{if } x < A \\ \frac{x-A}{B-a} & < \text{if } x \in [A, B] \\ 1 & \text{if } x \geq B \end{cases}$$

Then, the expected value is

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x \cdot f(x)dx = \int_A^B \frac{x}{B_A}dx \\ &= \frac{B+A}{2} \end{aligned}$$

The variance is

$$\begin{aligned} V(X) &= \int_{-\infty}^{\infty} (x - \mu_x)^2 f_X(x)dx = \int_A^B \frac{\left(x - \frac{B+A}{2}\right)^2}{B_A}dx \\ &= \frac{(B+A)^2}{12} \end{aligned}$$

If $a, b \in [A, B]$ and $a < b$, then

$$P(a \leq X \leq b) = \int_a^b \frac{x}{B-a} = \frac{b-a}{B-A}$$

Normal Distribution

Definition 4.1
We define **Normal Distribution** using the following two parameters

$$\begin{aligned} M &\rightarrow \text{ mean} \\ \sigma^2 &\rightarrow \text{ variance} \end{aligned}$$

denoted as $N(M, \sigma^2)$

Note: $\sigma = \sqrt{\sigma^2}$ is the standard deviation.

Definition 4.2
Suppose X is normally distributed, i.e $X \sim N(M, \sigma^2)$. Then, the probability density function of X is defined as

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{\frac{-(x-M)^2}{2\sigma^2}}, \; x \in (-\infty, \infty)$$

And the cumulative density function is

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} e^{\frac{-(t-M)^2}{2\sigma^2}} dt$$

Therefore,

$$P(a \leq X \leq b) = \frac{1}{\sqrt{2\pi}\sigma} \int_{a}^{b} e^{\frac{-(t-M)^2}{2\sigma^2}} dt$$

Note:

1. If $x \sim N(M, \sigma^2)$, then

$$\begin{aligned} E(X) &= M \\ V(X) &= \sigma^2 \end{aligned}$$

2. When $M = 0, \sigma^2 = 1$, it is a **Standard Normal Distribution** $Z \sim N(0,1)$, and the standard normal distribution probability density function is

$$f(z) = \frac{1}{\sqrt{2\pi}} \cdot e^{\frac{-z^2}{2}}, \; x \in (-\infty, \infty)$$

and the cumulative density function of Z is denoted as $\Phi(z)$,

$$P(Z \leq z) = \Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{\frac{-t^2}{2}} dt$$

To be specific about parameters, we denote probability density function of $X \sim N(M, \sigma^2)$ as

$$f(x; M, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{\frac{-(x-M)^2}{2\sigma^2}}, \; x \in (-\infty, \infty)$$

3. If $X \sim N(M, \sigma^2)$ and $Z \sim N(0,1)$, then

$$\begin{aligned} P(X \leq x) &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} e^{\frac{-(t-M)^2}{2\sigma^2}} dt \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x-M}{\sigma}} e^{\frac{-s^2}{2}} ds && \text{Change of vaiable with } s = \frac{t-M}{\sigma} \\ &= P(z \leq \frac{x-M}{\sigma}) \end{aligned}$$

Therefore, if $X \sim N(M, \sigma^2)$, then $\frac{X-M}{\sigma}$ has the standard normal distribution.

4. If $X \sim N(M, \sigma^2)$, then $Z = \frac{X-M}{\sigma} \sim N(0,1)$. Then

$$\begin{aligned} P(a \leq X \leq b) &= P\left(\frac{a-M}{\sigma} \leq z \leq \frac{b-M}{\sigma}\right) \\ &= \Phi\left(\frac{b-M}{\sigma}\right) - \Phi\left(\frac{a-M}{\sigma}\right) \end{aligned}$$

Therefore, we can use the standard normal cumulative density function to calculate probability for the random variable $X \sim N(M, \sigma^2)$, i.e

$$\begin{aligned} X \sim N(M, \sigma^2) &\longrightarrow Z \sim N(0,1) \\ x &\mapsto Z = \frac{x-M}{\sigma} && \text{It is also Z-score of } x \\ z &\mapsto x = Z \cdot \sigma + M \end{aligned}$$

4.3.1 Percentile and Critical Values

Definition 4.3
Suppose $Z \sim N(0,1)$, the standard normal distribution. Then, for $\alpha \in (0,1)$, z_α is the αth **critical value** that satisfies the following

$$\begin{aligned} P(Z \geq z_\alpha) &= \alpha \\ \text{i.e} \qquad \int_{z_\alpha}^{\infty} f(z;0,1)dz &= \alpha \\ \text{i.e} \qquad 1 - P(Z \leq z_\alpha) = \alpha &\implies 1 - \Phi(z_\alpha) = \alpha \end{aligned}$$

Definition 4.4
Let $p \in (0,1)$. Then $(100p)th$ **percentile**, $\eta(p)$ satisfies

$$\begin{aligned} P(Z \leq \eta(p)) &= p \\ \text{i.e} \quad \Phi(\eta(p)) &= p \end{aligned}$$

Suppose $X \sim N(M, \sigma^2)$ and $\alpha \in (0,1)$, then

$$\alpha^{th} \text{ critical value for } X := x_\alpha = \sigma z_\alpha + M$$

where z_α is the α^{th} critical value for $Z \in N(0,1)$

4.3.2 Approximating Bin(n,p) using Normal Distribution

Let $X \sim Bin(n,p)$, i.e $\mu_x = np$ and $\sigma_x^2 = np(1-p)$. Then if the $Bin(n,p)$ is not too skewed, then

$$P(X \leq x) \approx \Phi\left(\frac{x + 0.5 - Mx}{\sigma_x}\right)$$

The right hand side does not involve terms of the form $\binom{n}{k}$, which can be difficult to compute.

Gamma Distribution

Gama distribution is useful when modeling component lifetimes and waiting times.

Definition 4.5
Gamma function is defined as

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} \cdot e^{-t} dt, \text{ for } \alpha > 0$$

Theorem 4.6
Properties of Gamma Function

1. $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$ for $\alpha > 0$. Since $\Gamma(1) = 1$, we have $\Gamma(n) = (n-1)!$ for $n \in \mathbb{N}$ using induction
2. $\Gamma(\frac{1}{2}) = \sqrt{\pi}$

Definition 4.7
We say X has the **Gamma Distribution** with shape parameter α and scale parameter β if the probability density function of X is

$$f(x;\alpha,\beta) = \begin{cases} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} \cdot e^{\frac{-x}{\beta}} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

Since $f(x;\alpha,\beta) \geq 0 \; \forall x \in \mathbb{R}$ and we can check $\int_{-\infty}^{\infty} f(x;\alpha,\beta)dx = 1$ by doing a change of variable $y = \frac{x}{\beta}$, therefore f satisfies the property of probability density function.

Note:

1. When $\beta = 1$, we call it a **standard gamma distribution** with shape parameters α, with probability density function

   $$f(x;\alpha) = \begin{cases} \frac{x^{\alpha-1}e^{-x}}{\Gamma(\alpha)} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

   If T has standard Gamma distribution with shape parameters: α, then $X = \beta T$ has Gamma distribution with shape α, and scale β

2. If $X \sim Gamma(\alpha, \beta)$, then

   (a) $E(X) = \mu_x = \alpha\beta$

   (b) $V(X) = \sigma_X^2 = \alpha\beta^2$

   (c) $\sigma = \sqrt{\alpha}\beta$

3. The cumulative density function of $X \sim Gamma(\alpha, \beta)$ is the following

$$F_X(x; \alpha, \beta) = \begin{cases} \int_0^x f(t; \alpha, \beta)dt & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

   Then $P(X \leq x) = F_X(x; \alpha, \beta) = F\left(\frac{x}{\beta}; \alpha\right)$, the cumulative density function of standard Gamma with parameters α, i.e

$$F_X(x; \alpha) = \begin{cases} \int_0^x \frac{t^{\alpha-1}e^{-t}}{\Gamma(\alpha)} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

4.4.1 Special Cases of Gamma Distribution

Definition 4.8
When X has a Gamma Distribution with $\alpha = 1, \Gamma - \frac{1}{\lambda}$ for some $\lambda > 0$. Then, we say X has a **Exponential Distribution** with parameter λ. The probability density function of X is the

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

And the cumulative density function of X

$$F_X(x; \lambda) = \begin{cases} 1 - e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Note:

1. If X has exponential distribution with parameters $\lambda > 0$, hen

   (a) $E(X) = \mu_x = \frac{1}{\lambda}$

   (b) $V(X) = \sigma_X^2 = \frac{1}{\lambda^2}$

   (c) $\sigma = \frac{1}{\lambda}$

2. If we have a Poisson process with rate α, then the exponential distribution with $\lambda = \alpha$ models the distribution of elapsed time between the occurrence of two successive events.

3. If $X \sim Exp(\lambda)$, then

$$\begin{aligned} P(X \geq t + t_0 \mid X \geq t_0) &= \frac{[\{X \geq t + t_0\} \cap \{X \geq t_0\}]}{P(X \geq t_0)} \\ &= \frac{P(X \geq t + t_0)}{P(X \geq t_0)} = \frac{1 - F(t + t_0; \lambda)}{1 - F(t_0; \lambda)} \\ &= e^{-\lambda t} = P(X \geq t) \end{aligned}$$

   This means that if X was modeling lifetime of a component, the distribution of additional lifetime is exactly the same as the original distribution of lifetime, i.e the exponential distribution has **meaningless property**

Definition 4.9
X is said to have **chi-squared** distribution with parameter υ (degress of freedom) if $X \sim Gamma(\alpha, \beta)$ with $\alpha = \frac{\upsilon}{2}, \beta = 2$. The probability density function of X is defined as the following

$$f(x; \upsilon) = \begin{cases} \frac{1}{2^{\frac{\upsilon}{2}} \cdot \Gamma(\frac{\upsilon}{2})} \cdot x^{\frac{\upsilon}{2}-1} \cdot e^{\frac{-x}{2}} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Note:

1. If X is a chi-squared distribution with parameters υ, then

   (a) $E(X) = \mu_x = \alpha \cdot \beta = \upsilon$

   (b) $V(X) = \sigma_X^2 = \alpha\beta^2 = 2\upsilon$

   (c) $\sigma = \sqrt{2\upsilon}$

2. Chi-square distribution plays important role in statistical inference.

3. If $X \sim N(M, \sigma^2)$, then $\left(\frac{X-M}{\sigma}\right)^2$ has chi-square distribution with $\upsilon = 1$

Lognormal Distribution

Definition 4.10
We say X has a **Lognormal Distribution** if $\ln(X)$ has normal distribution with parameters M and σ^2. i.e

$$\ln(X) \sim N(M, \sigma^2)$$

If X has lognormal distribution, then the probability density function is

$$f(x; M, \sigma) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma x} \cdot e^{\frac{-(\ln(x)-M)^2}{2\sigma^2}} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Note:

1. If $\sim N(M, \sigma^2)$ then
   (a) $E(\ln(X)) = M$
   (b) $V(\ln(X)) = \sigma^2$
2. If X has lognormal distribution, and $\ln(x) \sim N(M, \sigma^2)$, then
   (a) $E(X) = \mu_x = e^{M+\frac{\sigma^2}{2}}$
   (b) $V(X) = \sigma_X^2 = e^{(2M+\sigma^2)} \cdot (e^{\sigma^2} - 1)$
3. We can use standard normal table to calculate the probabilities

$$\begin{aligned} F_X(x; M, \sigma) &= P(X \leq x) = P(\ln(x) \leq \ln(x)) \\ &= P(Z \leq \frac{\ln(x) - M}{\sigma}) \qquad \because \ln(x) \sim N(M, \sigma^2) \\ &= \Phi\left(\frac{\ln(x) - M}{\sigma}\right) \qquad \text{for } x > 0 \end{aligned}$$

Beta Distribution

Beta distribution takes values in a finite interval, which is good to model proportions that naturally lie in $(0, 1)$

Definition 4.11
X is said to have **Beta Disribution** with parameters:$\alpha, \beta > 0$ if the probability density function is

$$f(x; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot x^{\alpha-1}(1 - x)^{\beta-1} \qquad x \in (0, 1)$$

Note:

1. We called Beta distribution be cause the Beta function is

$$B(\alpha, \beta) = \frac{\Gamma(\alpha) \cdot \Gamma(B)}{\Gamma(\alpha + \beta)} \qquad \text{for } \alpha, \beta > 0$$

   So that

$$f(x; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} \cdot x^{\alpha-1}(1 - x)^{\beta-1} \qquad x \in (0, 1)$$

2. If $X \sim Beta(\alpha, \beta)$, then
   (a) $E(X) = \mu_x = \frac{\alpha}{\alpha+\beta}$
   (b) $V(X) = \sigma_X^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

   Note that both μ_x and σ_X^2 are in $(0, 1)$
3. Depending on values of α, β, the probability density function $f(x; \alpha, \beta)$ has different shapes.
   (a) If $\alpha > 1, \beta = 1$, then it is strictly increasing
   (b) If $\alpha = 1, \beta > 1$, then it is strictly decreasing
   (c) If $\alpha < 1, \beta < 1$, then it is U-shaped
   (d) If $\alpha = \beta$, then it is symmetric about $\frac{1}{2}$ with $\mu_x = \frac{1}{2}$ and $\sigma_X^2 = \frac{1}{4(2\alpha+1)}$
   (e) If $\alpha = \beta = 1$, then it is a uniform distribution on $(0, 1)$

Cauchy Distribution

Definition 4.12
X is said to have **Cauchy Distribution** with parameter θ if the probability density function of X is

$$\begin{aligned} f(x;\theta) = \frac{1}{\pi} \cdot \frac{1}{1+(x-\theta)^2} & x \in (-\infty, \infty) \\ & \theta \in (-\infty, \infty) \end{aligned}$$

Note:

1. $E(X)$ and $V(X)$ do not exists if X has Cauchy distribution since

   $$\begin{aligned} E(X) = \int_{-\infty}^{\infty} xf(x)dx &= \int_{-\infty}^{\infty} \frac{1}{\pi} \cdot \frac{x}{1+(x-\theta)^2} dx \\ &= 0 \text{ since } \frac{x}{1+(x-\theta)^2} \text{ is an odd function} \end{aligned}$$

   This is not the case.

2. The graph of $f(x;\theta)$ is bell shaped like the Normal density but has heavier tails than Normal density.

5 Joint Distribution

Many experiment or applications require that we use more than one characteristic or variable related to the experimental unit. For example:

1. Studying Body Mass Index (BMI) need the height and the weight
2. Predicating house prices need many factors: location, house area, number of bedroom and bathrooms, etc.
3. Predicting the price of a stock depends on many performance metrics - perhaps how Elon tweets

In all these cases, we nee to calculate the probability of two or more events simultaneously, therefore, we want the **Joint Distribution** of variable of interest

Two Discrete Random Variable

Definition 5.1
Suppose X, Y are two discrete random variables. Then $\mathcal{X} :=$ values of X and $\mathcal{Y} :=$ values of Y. Then, the cartesian product of $\mathcal{X}, \mathcal{Y}$

$$\mathcal{X} \times \mathcal{Y} = \{(x, y) : x \in \mathcal{X}, y \in \mathcal{Y}\}$$

is the **Joint Sample Space**.
The **Joint Probability Mass Function**, i.e Joint pmf of X and Y is

$$P(x, y) := P(X = x \text{ and } Y = y)$$

Which is $P($ X takes value x and Y takes value y)

Note:

1. $p(x, y)$ satisfies
   (a) $p(x, y) \geq 0 \ \forall (x, y) \in \mathcal{X} \times \mathcal{Y}$
   (b) $\sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} = 1$
2. If $A \subseteq \mathcal{X} \times \mathcal{Y}$, i.e A is a set of events in the joint distribution, then

$$P(A) = \sum_{(x,y) \in A} p(x, y)$$

Definition 5.2
Given a joint sample space of X, Y, we define the **Marginal Probability Mass Function of X** as

$$p_X(x) := \sum_{y \in \mathcal{Y}} p(x, y) \text{ for fixed } x \in \mathcal{X}$$

Marginal Probability Mass Function of Y as

$$p_Y(y) := \sum_{x \in \mathcal{X}} p(x, y) \text{ for fixed } y \in \mathcal{Y}$$

Definition 5.3
Given a joint sample space of X, Y, we define the **conditional distribution function** of X given that $Y = y$ as

$$\begin{aligned} p_{X|Y} &:= P(X = x \text{ given that } Y = y) \\ &:= \frac{p(x, y)}{p_Y(y)} \end{aligned}$$

conditional distribution function of Y given that $X = x$

$$\begin{aligned} p_{Y|X} &:= P(Y = y\text{given that } X = x) \\ &:= \frac{p(x, y)}{p_X(x)} \end{aligned}$$

Important Identity:

$$p(x, y) = p_{X|Y}(x \mid y) \cdot p_Y(y) = p_{Y|X}(y \mid x) \cdot p_X(y)$$

Two Continuous Random Variables

Definition 5.4
Let X, Y be two continuous random variables. Then, the **Joint Sample Space** is given by

$$\mathbb{R} \times \mathbb{R} = \mathbb{R}^2$$

Then, the **joint probability density function**, i.e joint pdf for the random variables X and Y is

$$f : \mathbb{R} \to \mathbb{R} \to \mathbb{R}$$

that satisfies the following

1. $f(x,y) \geq 0 \; \forall (x,y) \in \mathbb{R}^2$
2. $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = 1$

Note:

1. If $A \subseteq \mathbb{R}^2$, i.e A is a set of events in the joint sample space, then

$$P((x,y) \in A) = \int_A \int f(x,y)dxdy$$

If $A = \{(x,y) : x \in [a,b], y \in [c,d]\}$ = a rectangle $[a,b] \times [c,d]$ Then

$$\begin{aligned} P((x,y) \in A) &= \int_a^b \int_c^d f(x,y)dxdy \\ &= P(a \leq X \leq b \text{ and } c \leq y \leq d) \end{aligned}$$

Definition 5.5
Given a joint sample space of X, Y, where X, Y are continious, we define the **Marginal Probability Mass Function for** X as

$$f_X(x) := \int_{-\infty}^{\infty} f(x,y)dy \text{ for } x \in (-\infty, \infty)$$

Marginal Probability Mass Function for Y as

$$f_Y(y) := \int_{-\infty}^{\infty} f(x,y)dx \text{ for } y \in (-\infty, \infty)$$

Definition 5.6
Given a joint sample space of X, Y,where X, Y are continious, we define the **conditional density function** of Y given that $X = x$ as

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)} \qquad y \in (-\infty, \infty)$$

conditional density function of X given that $Y = y$

$$f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)} \qquad x \in (-\infty, \infty)$$

Important Identity:

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

Independent Random Variables

Definition 5.7
We say two discrete variables X and Y are **independent** if

$$p(x,y) = p_X(x) \cdot p_Y(y) \qquad \forall (x,y) \in \mathcal{X} \times \mathcal{Y}$$

otherwise, we say X and Y are **dependent**

Definition 5.8
We say two continious variables X and Y are **independent** if

$$f(x,y) = f_X(x) \cdot f_Y(y) \qquad \forall (x,y) \in \mathbb{R} \times \mathbb{R}$$

otherwise, we say X and Y are **dependent**

Note: When X and Y are independent, we can calculate the joint probability density function by multiplying the probability density functions of X and Y

Expected Values and Variances of Joint Distribution

Suppose X, Y are two random variables that are either both discrete or both continuous. Then, define

$$\begin{aligned} p(x,y) &:= \text{ joint probability mass function of } X,Y \text{ if both discrete} \\ f(x,y) &:= \text{ joint probability density function of } X,Y \text{ if both continious} \end{aligned}$$

Definition 5.9
Then, if $h(x,y)$ is a function of X and Y, then we can define the **expected value** of $h(x,y)$ as

$$E(h(x,y)) = \begin{cases} \sum\limits_{(x,y)\in\mathfrak{X}\times\mathfrak{Y}} h(x,y)\cdot p(x,y) & \text{if } X,Y \text{are discrete} \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x,y)\cdot p(x,y) & \text{if } X,Y \text{are discrete} \end{cases}$$

Let

$$\begin{aligned} \mu_x &:= E(X) \qquad \text{Expected Value of } X \\ \mu_y &:= E(Y) \qquad \text{Expected Value of } Y \end{aligned}$$

Then, we can define

$$\begin{aligned} h(X,Y) &= (X-\mu_x)(Y-\mu_y) \\ &= \text{ Product of the deviation of X and Y} \\ &\quad \text{ from their respective rexpected values} \end{aligned}$$

Note:

1. If large values of X are associated with large values of Y, i.e $(X-\mu_x)\geq 0 \implies (Y-\mu_y)\geq 0$ and small values of X are associated with small values of Y, i.e $(X-\mu_x)\leq 0 \implies (Y-\mu_y)\leq 0$ Then, we have
$$h(X,Y) = (X-\mu_x)(Y-\mu_y)\geq 0$$
as $(X-\mu_x)$ and $(Y-\mu_y)$ are both positive or both negative.
2. If large X values are assosiated with small values, i.e $(X-\mu_x)\geq 0 \implies (Y-\mu_y)\leq 0$ and large values of Y are associated with small values of X, i.e $(Y-Y_X)\geq 0 \implies (X-\mu_x)\leq 0$. Then, we have
$$h(X,Y) = (X-\mu_x)(Y-\mu_y)\leq 0$$
as $(X-\mu_x)$ and $(Y-\mu_y)$ has opposite sign

Then the expected value of $h(x,y) = (X-\mu_x)(Y-\mu_y)$, is a good candidate to measure relationship between X and Y.

Definition 5.10
We define the **covariance** of X and Y as

$$\begin{aligned} Cov(X,Y) &= E[(X-\mu_x)(Y-\mu_y)] = E(XY) - E(X)\cdot E(Y) \\ &= \text{ Expected value of the product of deviation} \\ &\quad \text{ of X from } \mu_x \text{ and } Y \text{ from } \mu_y \end{aligned}$$

If X, Y are both discrete, then

$$Cov(X,Y) = \sum_x\sum_y (x-\mu_x)(y-\mu_y)p(x,y)$$

If X, Y are both continuous, then

$$Cov(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_x)(y-\mu_y)f(x,y)dxdy$$

Note:

1. If $X = Y$, then $Cov(X,Y) = Cov(X,X) = V(X)$
2. If X and Y have strong positive relationship, i.e large values of X associated with large values of Y and small values of X associated with small values of Y, then $Cov(X,Y)$ will be positive
3. If $Cov(X,Y)$ is positive, then we can say X and Y are positively associated.
Similarly, if $Cov(X,Y)$ is negative, then we can say X and Y are negatively associated.

However, covariance is not a good measure of association. Depends on the unit in which X and Y are measured, it can be made arbitrarily large or small by changing units.

Correlation Coefficient

Let $\sigma_X = \sqrt{V(X)}$ and $\sigma_Y = \sqrt{V(y)}$, we can make the $Cov(X,Y)$ by dividing $\sigma_X \cdot \sigma_Y$.

Definition 5.11
The **correlation** of X and Y is

$$\rho_{X,Y} = Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y}$$

Note:

1. $\rho_{X,Y}$ does not depend on the units of X and Y
2. $\rho_{X,Y}$ satisfies $-1 \le \rho_{X,Y} \le 1$
3. If a, c are both positive or both negative, then

   $$Corr(aX+b, cY+d) = Corr(X,Y)$$

   In praticular, if $Y = aX+b$, then

   $$\begin{aligned} Corr(X,Y) &= Corr(X, aX+b) = sign(a) \cdot Corr(X,X) \\ &= sign(a) = \pm 1 \end{aligned}$$

   Therefore, $Corr(X,Y)$ measures the degree of linear relationship between X and Y.

   If $\rho_{X,Y}$ close to 1, then there is a strong positive linear association between X and Y
   If $\rho_{X,Y}$ close to -1, then there is a strong negative linear association between X and Y
   If $\rho_{X,Y} = 0$, then X, Y are uncorrelated, i.e no linear relationship. Note that it does not mean that there is no relationship between X, Y
4. If X, Y are independent, then

   $$\begin{aligned} Cov(X,Y) &= 0 \\ Corr(X,Y) &= 0 \end{aligned}$$

   However, $Corr(X,Y) = 0$ does not necessarily imply that X and Y are independent.
5. Correlation does not imply causation. It only implies association between variables. It is possible that there is an underlying variable that is causing the positive association
6. Note that

   $$\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y} \implies Cov(X,Y) = \rho_{X,Y} \cdot \sigma_X \cdot \sigma_Y$$

Linear Combination of Random Variables

Definition 5.12
Suppose $X_1, X_2, X_3, \cdots, X_n$ are $n-$random variables such that

(i) $E(X_i) = M_i$, $i = 1, 2, 3, \cdots, n$

(ii) $V(X_i) = \sigma_i^2$, $i = 1, 2, 3, \cdots, n$

Let $a_1, a_2, \cdots, a_n \in \mathbb{R}$, we can define the **Linear Combination** of $X_1, \cdots, X_n$ as

$$Y = a_1X_1 + a_2X_2 + \cdots + a_nX_n$$

Theorem 5.13
Given $Y = a_1X_1 + a_2X_2 + \cdots + a_nX_n$ for $X_1, X_2, X_3, \cdots, X_n$ are $n-$random variables and $a_1, a_2, \cdots, a_n \in \mathbb{R}$,then

1. $$\begin{aligned} E(Y) &= E(a_1X_1 + a_2X_2 + \cdots + a_nX_n) \\ &= a_1M_1 + a_2M_2 + \cdots + a_nM_n \end{aligned}$$

2. $$\begin{aligned} V(Y) &= V(a_1X_1 + a_2X_2 + \cdots + a_nX_n) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_1 a_j \cdot Cov(X_i, X_j) \end{aligned}$$

In particular, if $X_1, X_2, X_3, \cdots, X_n$ are independent, then

$$\begin{aligned} V(Y) &= a_1^2 V(X_1) + a_2^2 V(X_2) + \cdots + a_n^2 V(X_n) \\ &= a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \cdots + a_n^2\sigma_n^2 \end{aligned}$$

Example: If $Z = aX + bY$, then

$$\begin{aligned} E(Z) &= a\mu_x + b\mu_y \\ V(Z) &= a^2\sigma_X^2 + b^2\sigma_X^2 + 2abCov(X, Y) \end{aligned}$$

If $a = 1, b = -1$, i.e $Z = X - Y$. Then

$$\begin{aligned} E(X - Y) &= \mu_x - \mu_y \\ V(X - Y) &= \sigma_X^2 + \sigma_Y^2 + 2\rho x, y \cdot \sigma_X \cdot \sigma_Y \end{aligned}$$

6 Random Sample and Statistics

Collecting data it is important aspect of doing statistic. We want to use a subset of the population to draw conclusion/inference about the entire population, and this subset is called a **sample**. The way a sample is obtained will affect our confidence in the conclusion we drive from using it.

Example 6.0.1
suppose we have a vaccine for Covid 19 and We would be interested in knowing its efficacy. Consider the following experiment:

1. First, get two samples of people, group **A** and group **B**
2. Then, give vaccine to group A and give placebo to group B.
3. Expose both groups to the virus, and measure the appropriate response variable
4. Apply statistical tests to check if there is statistically significant difference between the group responses

To get correct conclusions about the vaccine, group A and B need to be carefully selected. Therefore, we need to consider the following questions

(i) Is it possible that certain people can never make it to the samples?

(ii) Does the sample contain only young individuals?

(iii) Are the samples representative of the population?

(iv) Are the sample size large enough?

And the statistical tools are usually based on principles of probability which need a precise notion of randomness of a sample.

Random Sample

Definition 6.1
We say $\{X_1, X_2, \cdots, X_n\}$, a set of random variable, is a **random sample** if

(i) X_i has the same probability distribution for $i = 1, 2, \cdots, n$

(ii) the set $\{X_1, X_2, \cdots, X_n\}$ is an independent set of random variables

Note:

1. Sometimes a random sample is also referred as a set of independent and identically distributed random variable
2. The common distribution of all the X_i's is called **population distribution**
3. $$\{X_1, X_2, X_3, X_n\} \xmapsto{\text{Sampling Procdedure}} \{x_1, x_2, \cdots, x_n\}$$
   note that $\{X_1, X_2, X_3, X_n\}$ is random sample, i.e a collection of random variables and $\{x_1, x_2, \cdots, x_n\}$ is the sample data that changes every time we run the sampling procedure.

Example 6.1.1
Suppose we are getting a random sample of size 3 from the population $\{1, 2, 3, 4, 5\}$. To get a random sample, we need to sample with replacement.
Let $\{Num1, Num2, Num3\}$ be a random sample, i.e collection of random variables. Since we are sampling with replacement, there are $5 \times 5 \times 5$ ways of getting sample data. so each run of sampling procedure gets one of the element in

$$\begin{Bmatrix} (1,1,1) & (1,1,2) & (1,2,1) & (2,1,1) & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & (5,5,4) & (5,5,5) \end{Bmatrix}$$

Statistics

We are often interested in some features or characteristics of the random sample.

Definition 6.2
A **statistic** is a quality calculated using a random sample, say $\{x_1, x_2, \cdots, x_n\}$. Since the calculation of the statistic involves random variables, it will also be a random variable

Note:

1. The value of the statistic changes as the sample data changes

2. A statistic is a random variable defined on sample data corresponding to a random sample. For example, if $\mathcal{X}$:= values of the random variables in the random sample, and sample data is $\{(x_1, x_2, \cdots, x_n) \mid x_i \in \mathcal{X}, \forall i = 1, 2, \cdots, n\}$. Since $\{X_1, X_2, \cdots, X_n\}$ is an independent set, the **joint distribution** for $X_1, x_2, \cdots, X_n$ is given as

$$p(x_1, x_2, x_3, \cdots, x_n) = p_X(x_1) \cdot p_X(x_2) \cdots p_X(x_n)$$

   where $p_X(x_i)$ is the probability mass function of the common distribution X, If X is continuous, then we can use probability density function $f_X(x)$ to define the joint probability density function of $X_1, X_2, \cdots, X)n$.

   Then, we can use the joint probability mass function or probability density function of $X_1, X_2, \cdots, X_n$ to calculate the probability distribution of the statistic of interest, i.e the **sampling distribution** of the statistic

Example 6.2.1
Recall the example that Population = $\{1, 2, 3, 4, 5\}$ and random sample = $\{Num1, Num2, Num3\}$. Then we can have the following different statistic

Name of the statistic	Definition of the statistic	Examples of evaluation on sample data
Sample Mean	$\overline{X} = \frac{Num_1, Num_2, Num_3}{3}$	$\overline{X}(1,1,1) = \frac{1+1+1}{3} = 1$ $\overline{X}(2,1,2) = \frac{2+1+2}{3} = \frac{5}{3}$
Sample Median	$\widetilde{X} = median\{Num_1, Num_2, Num_3\}$	$\widetilde{X}(1,1,1) = 1$ $\widetilde{X}(2,1,4) = 2$
Sample Maximum	$\text{Max} = \max\{Num_1, Num_2, Num_3\}$	$\text{Max}(1,1,1) = 1$ $\text{Max}(1,2,5) = 5$
Sample Total	$T_o = Num_1 + Num_2 + Num_3$	$T_o(1,1,1) = 1+1+1 = 3$ $T_0(1,3,5) = 1+3+5 = 9$
Sample variance	$S^2 = \frac{(Num_1 - \overline{X})^2 + (Num_2 - \overline{X})^2 + (Num_3 - \overline{X})^2)}{2}$	$S^2(1,1,1) = \frac{(1-\overline{X}(1,1,1))^2 + (1-\overline{X}(1,1,1))^2 + (1-\overline{X}(1,1,1))^2}{2}$ $S^2(1,1,1) = 0$ $S^2(1,2,3) = \frac{(1-\overline{X}(1,2,3))^2 + (2-\overline{X}(1,2,3))^2 + (3-\overline{X}(1,2,3))^2}{2}$ $S^2(1,2,3) = \frac{(1-2)^2 + (2-2)^2 + (3-2)^2}{2} = 1$

Sampling Distribution and Central Limit Theorem

6.3.1 Sampling Distribution

Definition 6.3
Let $\{X_1, X_2, \cdots, X_n\}$ be a random sample size n where

(i) X_i's have the same distribution for $i = 1, 2, 3, \cdots, n$

(ii) The set $\{X_1, X_2, \cdots, X_n\}$ is an independent set of random variables

Let $\mathcal{X} :=$ values of the random variables in the random sample, call it

$$\mathcal{X} = \{x_1, x_2, x_3, \cdots, x_n\}$$

Then, the **joint sample space** for the random variables $X_1, X_2, \cdots, X_n$ is the n-Cartesian product of $\mathcal{X}$, i.e

$$\mathcal{X}^n = \{(x_1, x_2, \cdots, x_n); x_i \in \mathcal{X}\}$$

Since $X_1, X_2, \cdots, X_n$ are independent random variables, the **joint probability distribution** of $X_1, X_2, X_3, \cdots, X_n$ is

1. $p(x_1, x_2, \cdots, x_n) = p_X(x_1) \cdot p_X(x_2) \cdots p_X(x_n)$ if $X_i's$ are discrete and $p_X(x)$ is the common probability mass function of all X_i's

2. $f(x_1, x_2, \cdots, x_n) = f_X(x_1) \cdot f_X(x_1) \cdots f_X(x_n)$ if X_i's are continuous and $f_X(x)$ is the common probability density function of all X_i's

For brevity, we will work with discrete random variables. But continuous variables case is similar

Definition 6.4
Let T be a statistic for the random sample $\{X_1, X_2, \cdots, X_n\}$, i.e a random variable calculated using random variables $X_1, X_2, \cdots, X_n$. Then, T is a function defined as

$$T : \mathcal{X}^n \longrightarrow \mathbb{R}$$

where $\mathcal{X}^n =$ joint sample space of $X_1, X_2, \cdots, X_n$

To calculate the sampling distribution of T, we need to use the joint distribution of the random variables $X_1, X_2, \cdots, X_n$. Let $p(x_1, x_2, \cdots, x_n)$ be the joint probability mass of $X_2, X_2, \cdots, X_n$ assuming the X_i's are discrete.
Then, we want the probability mass function of T, i.e for $t \in \mathcal{T}$, we want $P(T = t)$ where $\mathcal{T} :=$ values of T

Definition 6.5
Let $E(t) = \{(x_1, x_2, \cdots, x_n) \in \mathcal{X}^n; T(x_1, x_2, x_3, \cdots, x_n) = t\}$. Then

$$\begin{aligned} P(T = t) &= \sum_{(x_1, x_2, \cdots, x_n) \in E(t)} p(x_1, x_2, \cdots, x_n) \\ &= \sum_{(x_1, x_2, \cdots, x_n) \in E(t)} p_X(x_1) p_X(x_2) \cdots p_X(x_n) \end{aligned}$$

where $p_X(x)$ is the common probability mass function of $X_1, X_2, \cdots, X_n$

Example 6.3.1
The Sample Mean and Sample Total:
Let $X = \{X_1, X_2, \cdots, X_n\}$ be random sample coming from a population with mean $= \mu$ and variance $= \sigma^2$, i.e

$$\begin{aligned} E(X_i) &= \mu \qquad \text{for } i = 1, 2, \cdots n \\ V(X_i) &= \sigma^2 \qquad \text{for } i = 1, 2, \cdots n \end{aligned}$$

The sample total for $\{X_1, X_2, \cdots, X_n\}$ is

$$T_o := X_1 + X_2 + \cdots + X_n$$

And the sample mean for $\{X_1, X_2, \cdots, X_n\}$ is

$$\begin{aligned} \overline{X} :&= \frac{X_1 + X_2 + \cdots + X_n}{n} \\ &= \frac{T_o}{n} \end{aligned}$$

Then, we can calculate

$$\begin{aligned} E(T_o) &= E(X_1 + X_2 + \cdots + X_n) \\ &= E(X_1) + E(X_2) + \cdots + E(X_n) \\ &= n\mu \end{aligned}$$

and

$$\begin{aligned} V(T_o) &= V(X_1 + \cdots + X_n) \\ &= V(X_1) + \cdots + V(X_n) \\ &= \sigma^2 + \cdots + \sigma^2 \\ &= n\sigma^2 \end{aligned}$$

Similarly, $E(\overline{X}) = \mu$ and $V(\overline{X}) = \frac{\sigma^2}{n}$

If we want to calculate the sampling distribution of $\overline{X}$ and T_o, we have to use Central Limit Theorem!

6.3.2 Central Limit Theorem

Theorem 6.6
Central Limit Theorem (Normal Case):
Suppose $\{X_1, \cdots, X_n\}$ is a random sample coming from a population with Normal distribution $N(\mu, \sigma^2)$. Then, for any $n \in \mathbb{N}$, the distribution of the sample mean $\overline{X}$ and the sample total T_o are

$$\begin{aligned} \overline{X} &\sim N\left(\mu, \frac{\sigma^2}{n}\right) \\ T_o &\sim N\left(n\mu, n\sigma^2\right) \end{aligned}$$

Similarly, the random variables

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \text{ and } \frac{T_o - n\mu}{\sqrt{n}\sigma}$$

have the standard normal distribution

Note: For Central Limit Theorem to hold for all $n \in \mathbb{N}$ (especially when n is small), we need the population to be normally distributed. However, we can drop normality assumption if n is large

Theorem 6.7
Central Limit Theorem (General Case):
If $\{X_1, X_2, \cdots, X_n\}$ is a random sample from a population with

$$\begin{aligned} \text{mean} &= \mu \\ \text{variance} &= \sigma^2 \end{aligned}$$

Then, if n is large enough, the distribution of $\overline{X}$ and T_o is approximately normal, i.e

$$\begin{aligned} \overline{X} &\sim N\left(\mu, \frac{\sigma^2}{n}\right) \\ T_o &\sim N\left(n\mu, n\sigma^2\right) \end{aligned}$$

the approximation improves as $n \to \infty$, i.e the distribution of the random variables

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \text{ and } \frac{T_o - n\mu}{\sqrt{n}\sigma}$$

approaches the standard normal distribution as $n \to \infty$

Example 6.3.2
Suppose the population $= \{0, 1\}$. The distribution of the population is the Bernoulli distribution, i.e

$$\begin{aligned} P(X = 1) &= P(\text{success}) = p = P_X(1) \\ P(X = 0) &= P(\text{failure}) = 1 - p = P_X(0) \end{aligned}$$

Let $\{X_1, X_2, \cdots, X_n\}$ be a random sample from $\{0, 1\}$. Then, we want to calculate the sampling distribution of T_o.
Suppose $n = 2$, then $\mathcal{X}^2 = \{(0, 0), (0, 1), (1, 0), (1, 1)\}$, and values of $T_o = \{0, 1, 2\}$. Also, $p(x, y) =$

$p_X(x) \cdot p_X(y)$ be the joint probability mass function of X_1 and X_2.
Then,

$$P(T_o = 0) = p(0,0) = p_X(0) \cdot p_X(0) = (1-p)^2$$

$$\begin{aligned} P(T_o = 1) &= p(1,0) + p(0,1) = p_X(0) \cdot p_X(1) + p_X(1) \cdot p_X(0) \\ &= (1-p)p + p(1-p) = 2p \cdot (1-p) \end{aligned}$$

$$P(T_o = 2) = p(1,1) = p_X(1) \cdot p_X(1) = p^2$$

Putting this into a table gives

t	0	1	2
$P(T_o = t)$	$(1-p)^2$	$2p(1-p)$	p^2

Suppose $n = 3$, now

$$\mathcal{X}^3 = \left\{ \begin{array}{cccc} (0,0,0) & (1,0,0) & (1,1,0) & (1,1,1) \\ & (0,1,0) & (1,0,1) & \\ & (0,0,1) & (0,1,1) & \end{array} \right\}$$

values of $T_o = \{0,1,2,3\}$
$p(x_1,x_2,x_3) = p_X(x_1) \cdot p_X(x_1) \cdot p_X(x_2)$

Now, er have

$$P(T_o = 0) = p(0,0,0) = p_X(0) \cdot p_X(0) \cdot p_X(0) = (1-p)^3$$

$$\begin{aligned} P(T_o = 1) &= p(1,0,0) + p(0,1,0) + p(0,0,1) \\ &= p_X(1) \cdot p_X(0) \cdot p_X(0) + p_X(0) \cdot p_X(1) \cdot p_X(0) + p_X(0) \cdot p_X(0) \cdot p_X(1) \\ &= p(1-p)(1-p) + (1-p)p(1-p) + (1-p)(1-p)p \\ &= 3(1-p)^2 p \end{aligned}$$

$$\begin{aligned} P(T_o = 2) &= p(1,1,0) + p(0,1,1) + p(1,0,1) \\ &= p_X(1) \cdot p_X(1) \cdot p_X(0) + p_X(0) \cdot p_X(1) \cdot p_X(1) + p_X(1) \cdot p_X(0) \cdot p_X(1) \\ &= 3p^2(1-p) \end{aligned}$$

$$P(T_o = 3) = p(1,1,1) = p_X(1) \cdot p_X(1) \cdot p_X(1) = p^3$$

Then the probability distribution table of T_o is

t	0	1	2	3
$P(T_o = t)$	$(1-p)^3$	$3p(1-p)^2$	$3(1-p)p^2$	p^3

Suppose $n = 4$, now,

$$\mathcal{X}^3 = \left\{ \begin{array}{ccccc} (0,0,0) & (1,0,0,0) & (1,1,0,0) & (1,1,1,0) & (1,1,1,1) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & (0,0,0,1) & (0,0,1,1) & (0,1,1,1) & \cdots \end{array} \right\}$$

And the joint distribution is

$$p(x_1,x_2,x_3,x_4) = p_X(x_1) \cdot p_X(x_2) \cdot p_X(x_3) \cdot p_X(x_4)$$

And the values of T_o is

$$T_o = \{0,1,2,3,4\}$$

And the distribution of T_o can be calculated as

t	0	1	2	3	4
$P(T_o = t)$	$(1-p)^4$	$4p(1-p)^3 p$	$4(1-p)^2 p^2$	$4(1-p)p^3$	p^4

In general for arbitrary n, values of $T_o = \{0, 1, 2, 3, \cdots, n\}$ and

$$P(T_o = k) = \binom{n}{k}(1-p)^{n-k} \cdot p^k \qquad k = 0, 1, 2, \cdots, n$$

Note that this is the same as probability mass function of the Binomial distribution with parameters n and p, Therefore,

$$T_o \sim \text{Bin}(n, p)$$

This make sense because to count the number of successes (1) for given sample of size n, where each sample entry is an outcome of independent Bernoulli trial is the Binomial Experiment. Then

$$\begin{aligned} T_o \sim \text{Bin}(n, p) \implies & E(T_o) = np \\ & V(T_0) = np(1-p) \end{aligned}$$

Also, the central limit theorem says for large enough n, $T_o \sim N(n\mu_X, n\sigma_X^2) = N(np, np \cdot (1-p))$ where $\mu_X = \text{mean}(X) = p$ and $\sigma_X^2 = \text{Var}(x) = p(1-p)$.
$T_o \sim \text{Bin}(n, p) \implies$ for large n, $N(np, np(1-p))$. This is Normal approximation for Binomial

7 | Point Estimation

Introduction to Point Estimation

One goal of statistic is to draw insights/inference about certain aspects of the population using sample data. For example, we might want to estimate the following

1. Average GPA of students on campus
2. Average time spent on recreational activities by students at UMD
3. Median age of everybody affiliated to UMD
4. The median yearly income of people in the USA

In each these situations, we need to identify

1. population of interest
2. a characteristic of the population that we are interested in

Definition 7.1
Suppose

$$\{X_1, X_2, X_3, \cdots, X_n\}$$

to be a random sample coming from a fixed population. Then, a **point estimator** for the population is any statistic $\widehat{\theta}$ for the random sample $\{X_1, X_2, \cdots, X_n\}$, i.e

$$\widehat{\theta} : \mathcal{X}^n \to \mathbb{R}, \qquad \mathcal{X} := \text{ values of } X_i$$

We typically expect the values of $\widehat{\theta}$ to be a sensible value of a certain population characteristic, which is a **population parameter** denoted by θ

Note:

1. The population parameter θ is fixed once we fix the population
2. If we have population data, i.e a census, then we can calculate the exact value of θ
3. Usually the population is intractable, which means we need to resort the sample data to get an estimate for θ
4. Suppose θ is a parameter of interest. Given an estimator $\widehat{\theta}$ for θ, $\widehat{\theta}$ is a statistic depending on the random sample $\{X_1, X_2, \cdots, X_n\}$. Therefore, the estimate $\widehat{\theta}$ will change everytime sample data changes

Definition 7.2
If $\widehat{\theta}$ is a point estimator for the population parameter θ, then the relation of average value of $\widehat{\theta}$ and true value θ, or **Bias of $\widehat{\theta}$** is

$$\text{Bias}(\widehat{\theta}) = E(\widehat{\theta} - \theta)$$

It is also the **expected error** when we use $\widehat{\theta}$ to estimate θ using a random sample of size n

Definition 7.3
We say $\widehat{\theta}$ is an **unbiased estimator** for θ if

$$\text{Bias}(\widehat{\theta}) = 0 \qquad \text{for all possible choices of } \theta$$

i.e $E(\widehat{\theta} - \theta) = 0$ for every possible θ

Exercise: Suppose we have estimators $\widehat{\theta}_1, \widehat{\theta}_2, \cdots, \widehat{\theta}_k$ estimating θ. Among all the $\widehat{\theta}_i$, is there a notion of one estimator being better than theothers?

Principle of Unbiased Estimation

Example 7.2.1
Let population be $\{0,1\}$ and the population distribution be Bernolli(p), i.e $P(X = 1) = p$. Let $\{X_1, X_2, \cdots, X_n\}$ be a random sample. Recall that

$$\begin{aligned} T_o &:= \text{ sample total} \\ &= X_1 + X_2 + \cdots + X_n \\ T_o &\sim \text{Bin}(n, p) \end{aligned}$$

Let's define

$$\hat{p} := \frac{T_o}{n}$$

This is the sample proportion of successes in a sample of size n. And $\widehat{p}$ is an estimator for p, the true probability of getting a success in a single Bernoulli trial.

Now,

$$E(\widehat{p}) = E\left(\frac{T_o}{n}\right) = \frac{E(T_o)}{n} = \frac{n \cdot p}{n} = p$$

Therefore, $\widehat{p}$ is an unbiased estimator for p

Example 7.2.2
Suppose $\{X_1, X_2, \cdots, X_n\}$ is a random sample from a population with mean μ and variance σ^2. Then, we define the following:

$$\begin{aligned} \widehat{\theta_1} &:= \overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} && \text{sample mean} \\ \widehat{\theta_2} &:= \overline{X} + X_1 - X_m && \text{sample mean plus 1st and second random sample} \end{aligned}$$

Then,

$$\begin{aligned} E(\widehat{\theta_1}) &:= E(\overline{X}) = \mu \\ E(\widehat{\theta_2}) &:= E(\overline{X} + X_1 - X_m) \\ &= \mu + \mu - \mu = \mu \end{aligned}$$

Therefore, both $\widehat{\theta_1}$ and $\widehat{\theta_2}$ are biased estimations for he population mean μ. Therefore, unbiased estimation does not guarantee a unique choice of estimator.

Definition 7.4
Suppose $\widehat{\theta}$ is a statistic for a random sample $\{X_1, X_2, \cdots, X_n\}$ such that

$$E(\widehat{\theta}) = \theta \qquad \text{for all possible choices of } \theta$$

Then, $\sigma^2_{\widehat{\theta}}$:= **variance of the sampling distribution of** $\widehat{\theta}$

Theorem 7.5
Principle of Minimum Variance Unbiased Estimation:
Among all unbaised estimators of θ, the one that has the minimum variance is called the **Minimum Variance Unbiased Estimator** of θ

Example 7.2.3
We know that $\widehat{\theta_1} = \overline{X}$ and $\widehat{\theta_2} = \overline{X} + X_1 - X_n$ are both unbiased estimators of μ
To rank them, we calculate $V(\widehat{\theta_1})$ and $V(\widehat{\theta_2})$.

$$\begin{aligned} V(\widehat{\theta_1}) &= V\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) \\ &= V\left(\frac{X_1}{n} + \frac{X_2}{n} + \cdots + \frac{X_n}{n}\right) \\ &= \frac{1}{n^2}\left(\sigma^2 + \sigma^2 + \cdots + \sigma^2\right) \\ &= \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \end{aligned}$$

Similarly, we can show that

$$\begin{aligned} V(\widehat{\theta_2}) &= V(\overline{X} + X_1 - X_n) \\ &= \frac{\sigma^2}{n} + 2\sigma^2 \\ &= \frac{2n+1}{n}\sigma^2 \end{aligned}$$

Since $\frac{2n+1}{n} > \frac{1}{n}$ for all $n \in \mathbb{N}$, therefore, $V(\widehat{\theta}_2) > V(\widehat{\theta}_1)$. Therefore, if we were to choose only between $\widehat{\theta}_1$ and $\widehat{\theta}_1$, we would go with $\widehat{\theta}_1$. However, this does not prove that $\overline{X}$ is the Minimum Variance Unbiased Estimator

Theorem 7.6
If $\{X_1, X_2, \cdots, X_n\}$ is a random sample from a normally distributed population, i.e $\mathrm{N}(\mu, \sigma^2)$. Then the estimator $\overline{X}$ is the Minimum Variance Unbiased Estimator for μ.
And the **standard error** of a point estimator $\widehat{\theta}$ is $\sigma_{\widehat{\theta}} = \sqrt{V(\widehat{\theta})}$, which provides a measure or precision of the point estimator $\widehat{\theta}$, i.e the standard deviation of the sampling distribution of $\widehat{\theta}$

Note:

1. If $\widehat{\theta}$ has a continuous distribution, then

$$\begin{aligned} P(\widehat{\theta} = \theta) &= P(\text{The estimator } \widehat{\theta} \text{ takes the value } \theta \text{ the true parameter value}) \\ &= 0 \end{aligned}$$

Even so, the point-estimator provides an exact estimate for θ, we have zero confidence that the calculated point estimate will equal θ

Methods of Point Estimation

Suppose Population has distribution X and probability mass function or probability density function of X is $f(x)$.
Let $\{X_1, X_2, \cdots, X_n\}$ be a random sample of size n from population.

Then for $k = 1, 2, 3, \cdots$

1. The kth population moment or the kth distribution moment is the expected values of the random variable X^k, i.e $E(X^k)$

2. The kth sample moment for the random sample $\{X_1, X_2, \cdots, X_n\}$ is

$$\frac{X_1^k + X_2^k + \cdots + X_n^k}{n} \qquad \text{i.e} \qquad \frac{\sum_{i=1}^n X_i^k}{n}$$

7.3.1 The Method of Moments

If there are m parameters: $\theta_1, \theta_2, \theta_3, \cdots, \theta_m$ tand we want to estimate using $\{X_1, X_2, \cdots, X_n\}$. It is the same to get m-equations by equating the first m population moments to the first m sample moments and pray we can solve these equations to get estimators $\widehat{\theta}_1, \widehat{\theta}_2, \cdots \widehat{\theta}_n$ for $\theta_1, \theta_2, \cdots, \theta_m$ respectively.

Example 7.3.1
Sampling From Exponential Distribution
Suppose $\{X_1, X_2, X_3, \cdots, X_n\}$ be a random sample from $\exp(\lambda)$. Note that there is only one parameter to estimate and therefore only need to calculate first moments.

Then, the first population moment is

$$E(X) = \frac{1}{\lambda}$$

and the first sample moment is

$$\frac{1}{n}\left(\sum_{i=1}^{n} X_i\right) = \overline{X}$$

, the sample mean. Then, equating the population moments to sample moments, we got

$$\frac{1}{\lambda} = \overline{X} \implies \lambda = \frac{1}{\overline{X}}$$

Therefore, method of moment estimator for λ is

$$\widehat{\lambda} = \frac{1}{\overline{X}}$$

Example 7.3.2
Sampling From Normal Distribution
Suppose $\{X_1, X_2, X_3, \cdots, X_n\}$ be a random sample from $N(\mu, \sigma^2)$. We want to estimate μ and σ^2 (i.e

$k = 2)$
First, calculate the kth population moments:

$$\begin{aligned} k = 1 &\to E(X) = \mu \\ k = 2 &\to E(X^2) \end{aligned}$$

Recall that the identity $\sigma^2 = V(X) = E(X^2) - (E(X))^2$, therefore, we got

$$E(X^2) = \sigma^2 + \mu^2$$

Then,equating the population moments to sample moments:

$$\begin{aligned} E(X) = \mu &= \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X} \\ E(X^2) = \sigma^2 + \mu^2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 \end{aligned}$$

Therefore, we got the moment estimators as the following

$$\begin{aligned} \widehat{\mu} &= \overline{X} \\ \widehat{\sigma^2} &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 \end{aligned}$$

Example 7.3.3
Sampling From Gamma Distribution
Suppose $\{X_1, X_2, X_3, \cdots, X_n\}$ be a random sample from Gamma(α, β), i.e α is the shape and β is the scale.
Recall that $E(X) = \alpha\beta$, $V(X) = \alpha\beta^2$. Since $E(X^2) - E(X)^2 = \alpha\beta^2$, then $E(X^2) = \alpha\beta^2 + (E(X))^2$. Therefore, we got the following population moments:

$$\begin{aligned} E(X) &= \alpha\beta \\ E(X^2) &= \alpha\beta^2 + (\alpha\beta)^2 \\ &= \alpha\beta^2(1+\alpha) \end{aligned}$$

Evaluating sample moments to population moments, we get

$$\begin{aligned} \overline{X} &= \alpha\beta \\ \frac{\sum_{i=1}^{n} X_i^2}{n} &= \alpha\beta^2(1+\alpha) \end{aligned}$$

Then let $A = \overline{X}$ and $B = \frac{1}{n}\sum_{i=1}^{n} X_i^2$, we need to solve the following for α, β

$$\begin{aligned} A &= \alpha\beta \\ B &= \alpha\beta^2(1+\alpha) \end{aligned}$$

set $\beta = \frac{A}{\alpha}$, then

$$\begin{aligned} B &= \alpha \cdot \left(\frac{A}{\alpha}\right)^2 (1+\alpha) \\ &= \frac{A^2(1+\alpha)}{\alpha} \end{aligned}$$

Therefore,

$$\begin{aligned} \alpha\beta &= A^2 + A^2\alpha \\ \implies \alpha(B - A^2) &= A^2 \\ \implies \alpha &= \frac{A^2}{(B - A^2)} \end{aligned}$$

Therefore, the method of moment estimators are

$$\begin{aligned} \widehat{\alpha} &= \frac{\overline{X}^2}{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2} \\ \widehat{\beta} &= \frac{\overline{X}}{\widehat{\alpha}} \end{aligned}$$

7.3.2 Maximum Likelihood Estimation

Definition 7.7
Suppose $\{X_1, X_2, \cdots, X_n\}$, a random sample from population with probability density function/probability mass function $f(x)$.
Let $f(x_1, x_2, \cdots, x_n)$ be the joint probability density function/probability mass function of $X_1, X_2, \cdots, X_n$.
Suppose X_i's are independent, then

$$f(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} f(x_i)$$

Now, given sample data $\{x_1, x_2, \cdots, x_n\}$, we want to estimate parameters $\theta_1, \theta_2, \cdots, \theta_n$. Then, the **Likehood Function** for sample data $\{x_1, x_2, \cdots, x_n\}$ is defined as

$$f(x_1, x_2, \cdots, x_n; \theta_1, \theta_2, \cdots, \theta_n)$$

And **Maximum Likelihood Estimation** is finding choices of parameters $\widehat{\theta_1}, \cdots, \widehat{\theta_n}$ which maximize the likehood function for the observed sample data $\{x_1, x_2, x_3, \cdots, x_n\}$. That is, find $\widehat{\theta_1}, \cdots, \widehat{\theta_n}$ such that

$$f(x_1, x_2, \cdots, x_n; \widehat{\sigma_1}, \widehat{\sigma_2}, \cdots, \widehat{\sigma_n}) \geq f(x_1, x_2, \cdots, x_n; \theta_1, \theta_2, \cdots, \theta_n)$$

for all possible values of $\theta_1, \theta_2, \cdots, \theta_n$.

Then, $\widehat{\theta_1}, \cdots, \widehat{\theta_n}$ is the **maximum likelihood estimators** for $\theta_1, \cdots, \theta_n$ respectively, using $\{x_1, x_2, \cdots, x_n\}$.
If we substitute X_i for x_i, we will get maximum likelihood estimators $\widehat{\theta_1}, \cdots, \widehat{\theta_n}$ for $\theta_1, \theta_2, \cdots, \theta_n$ respectively.

Theorem 7.8
Properties of Maximum Likelihood Estimators:

1. **Invariance Principle**: If $\widehat{\sigma_1}, \widehat{\sigma_2}, \cdots, \widehat{\sigma_n}$ are maximum likelihood estimators for $\sigma_1, \cdots, \sigma_n$, then $h(\widehat{\sigma_1}, \cdots, \widehat{\sigma_n})$ is an maximum likelihood estimators for $h(\sigma_1, \cdots, \sigma_n)$

2. **Large Sample Behavior**: When sample size is large, the maximum likelihood estimators $\widehat{\sigma}$ for σ is atleast approximate unbiased, i.e $\widehat{\sigma} \approx \sigma$ and has variance that is either small as (or nearly as small as) can be achieved by any estimator.

Example 7.3.4
Sampling From Exponential Distribution:
Suppose $\{X_1, X_2, X_3, \cdots, X_n\}$ be a random sample from $\exp(\lambda)$. Then the probability density function of X_i is $f(x_i; \lambda) = \lambda e^{-\lambda x_i}$.

Then the likelihood function is

$$\begin{aligned} f(x_1, x_2, \cdots, x_n; \lambda) &= \prod_{i=1}^{n} f(x_i; \lambda) \\ &= \lambda e^{-\lambda x_1} \cdot \lambda e^{-\lambda x_2} \cdots \lambda e^{-\lambda x_n} \\ &= \lambda^n e^{-\lambda(x_1 + x_2 + \cdots + x_n)} \end{aligned}$$

Therefore,

$$f(x_1, x_2, \cdots, x_n; \lambda) = \lambda^n \ \exp(-\lambda \sum_{i=1}^{n} x_i)$$

Then, we want to maximize this as a function of λ's.
If $g(\lambda) = \lambda^n \ \exp(-\lambda \sum_{i=1}^{n} x_i)$, then

$$\begin{aligned} \ln(g(\lambda)) &= \ln(\lambda^n \ \exp(-\lambda \sum_{i=1}^{n} x_i)) \\ &= \ln(\lambda^n) + \ln(\exp(-\lambda \sum_{i=1}^{n} x_i)) \\ &= n \ln(\lambda) - \lambda \sum_{i=1}^{n} x_i \end{aligned}$$

This is log likelihood function, which is easier to maximize.
Let $h(\lambda) = \ln(g(\lambda))$, then

$$\begin{aligned} h'(\lambda) &= \frac{d}{d\lambda}\left(n\ln(\lambda) - \lambda\sum_{i=1}^{n} x_i\right) \\ &= n \cdot \frac{1}{\lambda} - \sum_{i=1}^{n} x_i \end{aligned}$$

Therefore,

$$h'(\lambda) = 0 \implies n \cdot \frac{1}{\lambda} = \sum_{i=1}^{n} x_i \implies \lambda = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} x_i}$$

Also, $h''(\lambda) = \frac{-n}{\lambda^2} < 0 \implies \lambda = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} x_i}$ is a point of local maxima.

Since ln is an increasing function, then $\lambda = \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^{-1}$ maximizes $h(\lambda) \implies \lambda = \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^{-1}$ maximizes the likelihood function $g(\lambda)$. Therefore, the maximum likelihood estimator for λ is $\widehat{\lambda} = \frac{1}{\overline{X}}$, and this is the same as the moment estimator for λ

Example 7.3.5
Sampling From Normal Distribution
Suppose $\{X_1, X_2, X_3, \cdots, X_n\}$ be a random sample from $N(\mu, \sigma^2)$. Then X_i has the following probability density function

$$f(x_i; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

Then, the likelihood function is given as

$$\begin{aligned} f(x_1, x_2, \cdots, x_n; \mu, \sigma^2) &= \prod_{i=1}^{n} f(x_i; \mu, \sigma^2) \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \cdot e^{-\sum_{i=1}^{n}\left(\frac{(x_i-\mu)^2}{2\sigma^2}\right)} \end{aligned}$$

So that the log-likelihood function is

$$\begin{aligned} \ln(f(x_1, x_2, \cdots, x_n; \mu, \sigma^2)) &= \ln\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n + \ln\left(e^{-\sum_{i=1}^{n}\left(\frac{(x_i-\mu)^2}{2\sigma^2}\right)}\right) \\ &= \frac{n}{2}\ln\left(\frac{1}{2\pi\sigma^2}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 \end{aligned}$$

Therefore,

$$h(\mu, \sigma^2) = \frac{n}{2}\ln\left(\frac{1}{2\pi\sigma^2}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

Then, we need to maximize $h(\mu, \sigma^2)$ to calculate critical values, i.e calculate critical value where $\frac{\partial h}{\partial \mu} = 0$ and $\frac{\partial h}{\partial \sigma^2} = 0$. We can use Hessian matrix to show that these are max.
Then, we can show that the maximum likehood estimators are

$$\begin{aligned} \widehat{\mu} &= \overline{X} \\ \widehat{\sigma^2} &= \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \end{aligned}$$

Note that while $\widehat{\mu}$ is unbiased, $\widehat{\sigma^2}$ is biased!