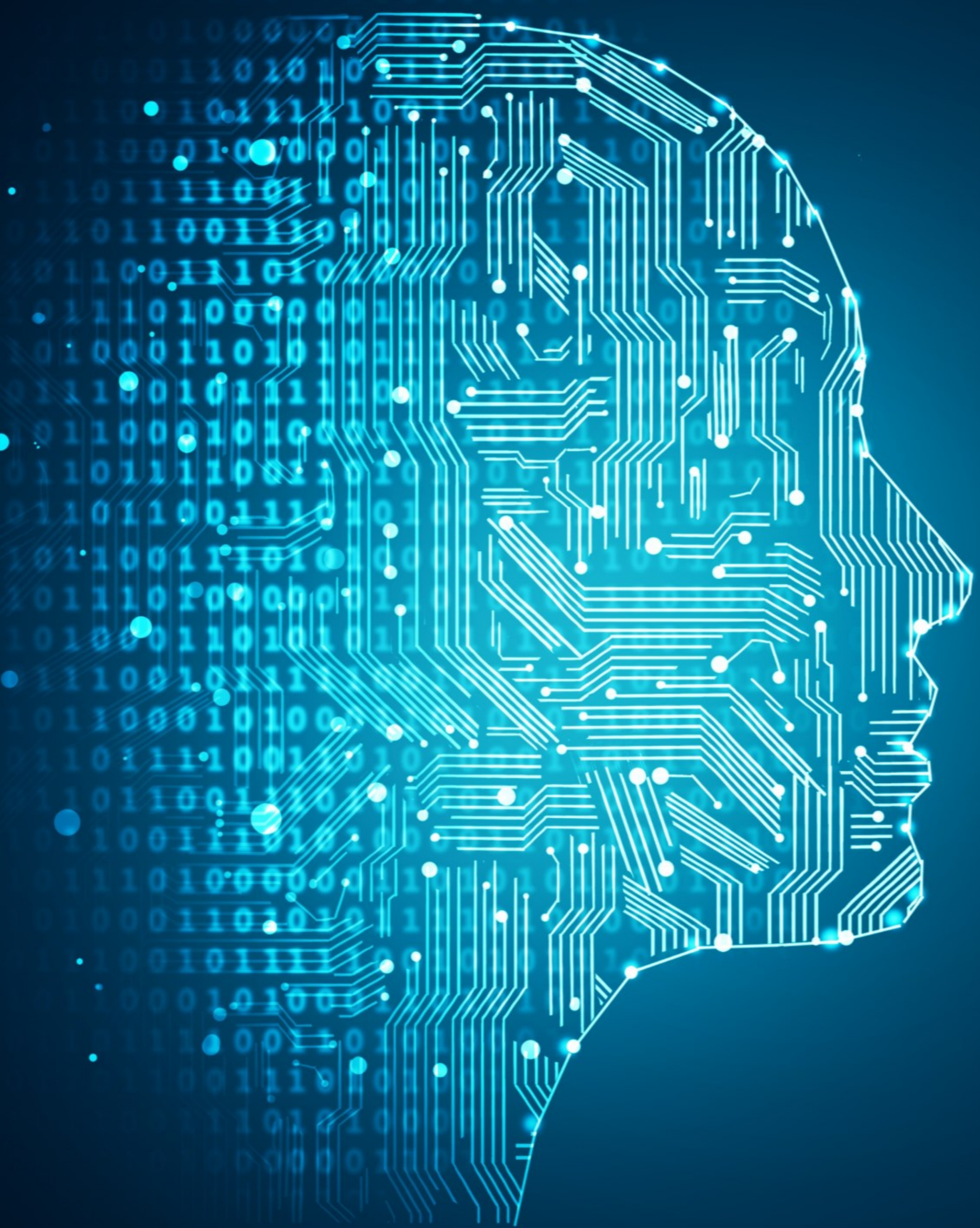


AI6103 Probability Theory

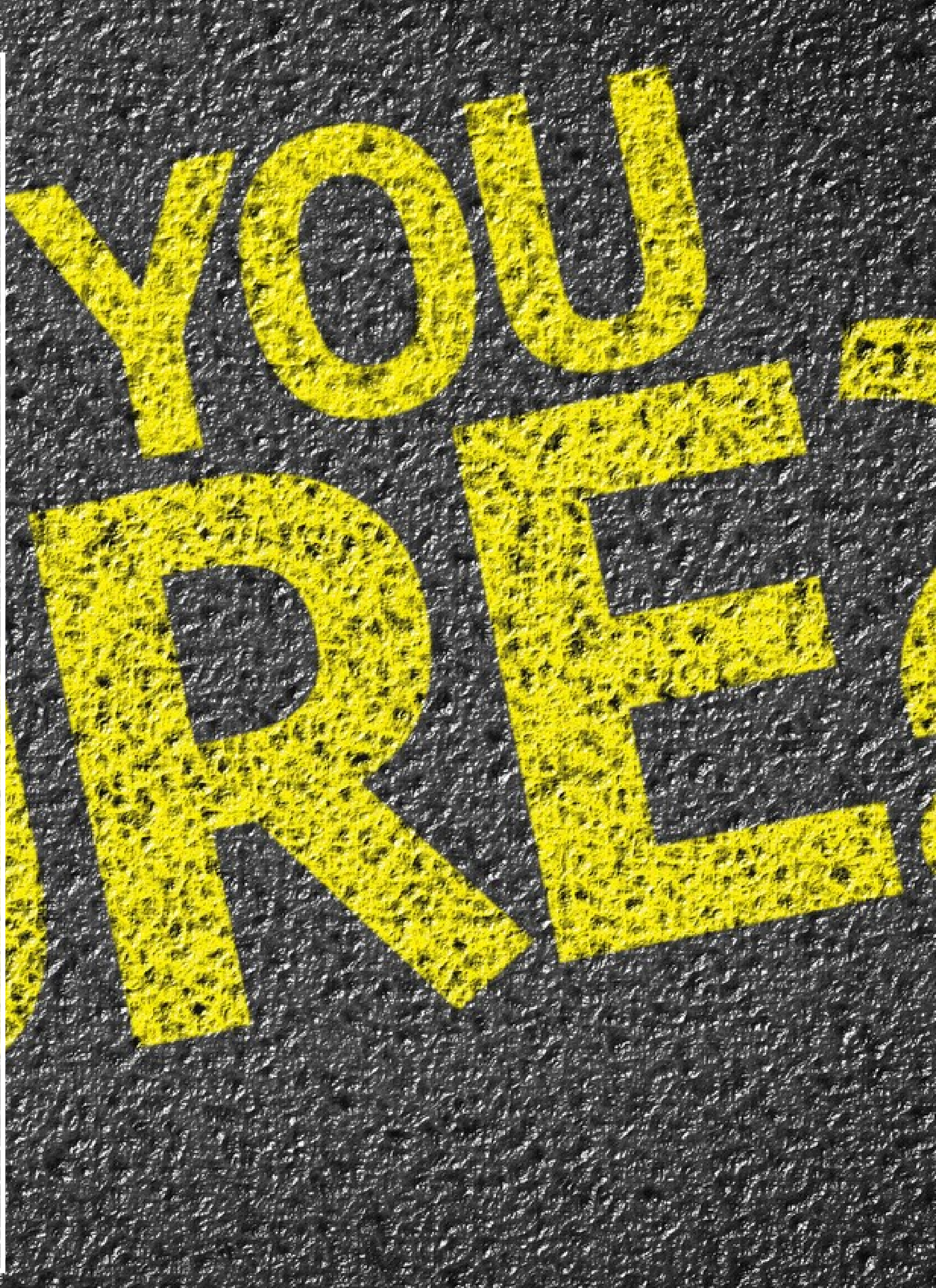
Boyang Li, Albert

School of Computer Science and Engineering

Nanyang Technological University

Logic Rules

- **IF** the time is 7am in Singapore, **THEN** the sun will rise in the next 10 minutes.
- **IF** the time is 8am in Singapore, **THEN** the NTU shuttle will arrive in the next 10 minutes.
- **IF** the PCR test result is positive, **THEN** the patient has COVID.
- **IF** the patient is having a fever and has high white cell counts, **THEN** the patient has an infection.

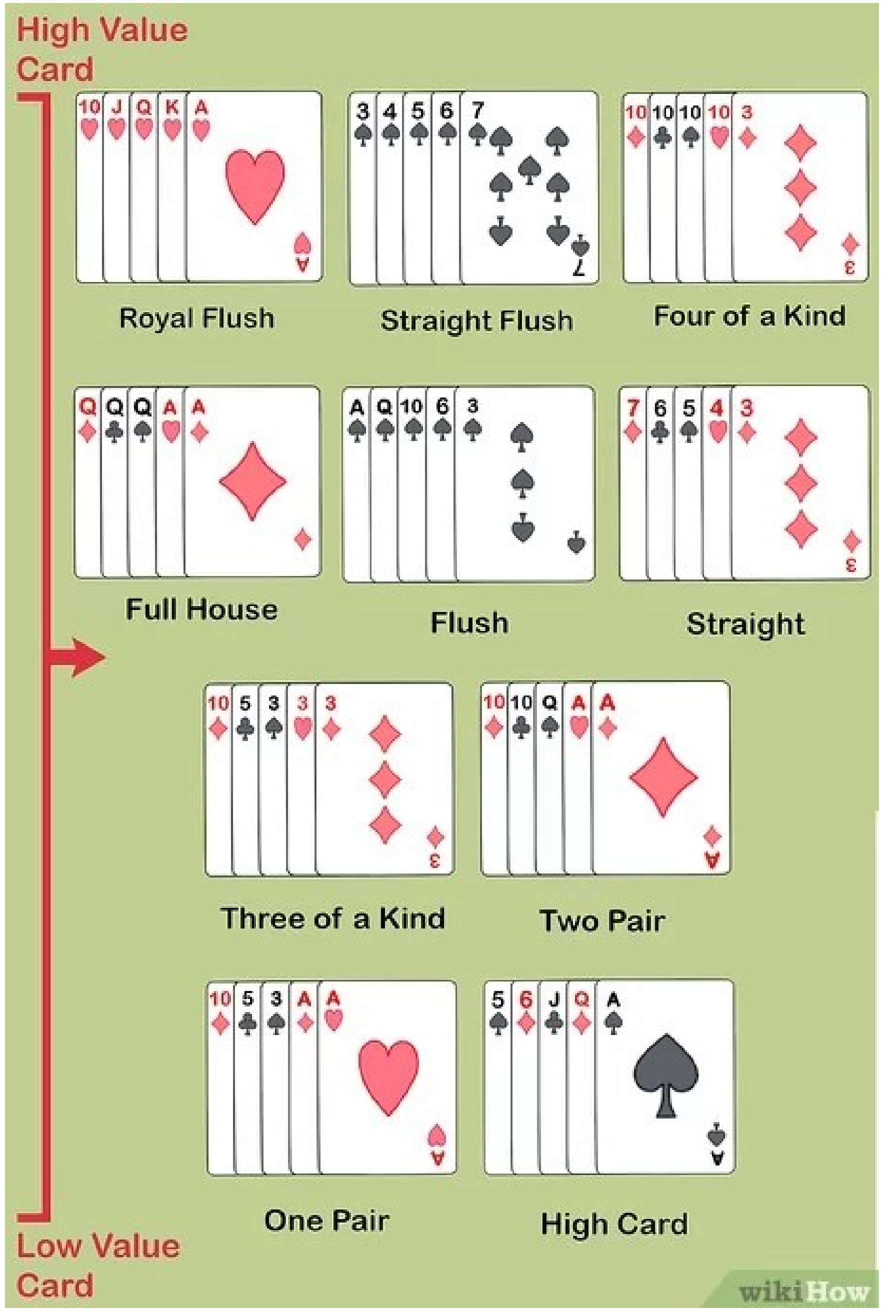


Tomatoes

Horses

Surprise, surprise

What are these?



How to design an AI that can win in these games?

The Need for Probability

- **Laziness**: It is too much work to list the complete set of exceptional cases like exceptions to a rule or extraordinary categories of images.
- **Theoretical ignorance**: We do not have a complete theory for most scientific and engineering domains. Thus, we cannot make predictions that are 100% accurate.
- **Practical ignorance**: Even if we know all the rules, we might be uncertain about a particular patient because we may not have enough resources (money, time, manpower, etc.) to run all tests.
- **Artificial ignorance**: In the case of games, the game rules prevent us from learning all facts such as the sequence of cards in a deck.

Materials and Textbooks

- Russel and Norvig. Artificial Intelligence: A Modern Approach, 4th ed.
  - Chapters 12-14.
- Goodfellow et al. Deep Learning
  - Chapter 3
- Available as e-books in NTU Library
  - Narayan C. Giri. Introduction to Probability and Statistics (rigor and good coverage)
  - Joel A. Nachlas. Probability Foundations for Engineers (easy to follow but incomplete)

Mathematics of a single six-sided die

- What are the possible outcomes?
- What are the probabilities of getting a 6?
- What are the probabilities of getting 1 <u>or</u> 6?
- What are the probabilities of getting 1 in the first trial and 6 in the second?

More formally …

This is not rigorous from a real analysis perspective, but we are going to simplify things a bit in this course.

- **Sample space** Ω: The set of possible outcomes ω of a random experiment
  - For a single dice roll: {1, 2, 3, 4, 5, 6}
  - For two consecutive rolls: {(1, 1), (1, 2), …, (6, 5), (6, 6)}
- **Event**: A subset of the sample space.
  - A single roll getting 6
  - Two rolls with the sum 11
- **Probabilistic model:** a function P that assigns a probability to every event.
- From the probabilistic model, we can compute the probability of events.

Three Axioms of Probability

- For any event A, its probability $\mathrm{P}(A) \geq 0$
- The probability of the entire sample space $\mathrm{P}(\Omega) = 1$
- If events A_1 and A_2 are disjoint, then
$$P(A_1 \cup A_2) = P(A_1) + P(A_2)$$

- From the above, we know
  - The probability of the empty set $\emptyset, \mathrm{P}(\emptyset)$ is 0.
  - For events A_1 and A_2, if $A_1 \subseteq A_2, \mathrm{P}(A_1) \leq \mathrm{P}(A_2)$
  - The complement of A in Ω, written as A^c, has the probability
$$P(A^c) = P(\Omega \backslash A) = 1 - P(A)$$

Three Axioms of Probability

- For any event A, its probability $\mathrm{P}(A) \geq 0$
- The probability of the entire sample space $\mathrm{P}(\Omega) = 1$
- If events A_1 and A_2 are disjoint, then
$$P(A_1 \cup A_2) = P(A_1) + P(A_2)$$
- From the above, we also know
  - The probability of every outcome $\omega \in \Omega$, $\mathrm{P}(\{\omega\})$ or simply $\mathrm{P}(\omega)$, is in $[0,1]$
  - If A_1 and A_2 are not disjoint, $P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$

Three Axioms of Probability

- For any event A, its probability $\mathrm{P}(A) \geq 0$
- The probability of the entire sample space $\mathrm{P}(\Omega) = 1$
- If events A_1 and A_2 are disjoint, then
  $$P(A_1 \cup A_2) = P(A_1) + P(A_2)$$

- For 6-sided die, $\Omega = \{1, 2, 3, 4, 5, 6\}$
- $P(\Omega) = P(\{1\} \cup \{2\} \cup \{3\} \cup \{4\} \cup \{5\} \cup \{6\})$
  $$= P(1) + P(2) + P(3) + P(4) + P(5) + P(6)$$
  $$= 1$$
- Thus, assuming every outcome is equally likely, we have
  $$P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = \frac{1}{6}$$

Exercise: Counting the outcomes

- Question: I have three letters and three recipients. I randomly put the letters into envelopes with the recipient's names, one for each envelope. What is the probability of at least one recipient receiving the correct letter?
- Letters: 1, 2, 3. Recipients: A, B, C.
- Total number of outcomes: 3x2=6 (Are they equally likely?)
- All outcomes: (A-1, B-2, C-3), (A-1, B-3, C-2), (A-2, B-1, C-3), (A-2, B-3, C-1), (A-3, B-1, C-2), (A-3, B-2, C-1)

Exercise: Counting the outcomes

- Question: I have three letters and three recipients. I randomly put the letters into envelopes with the recipient's names. What is the probability of at least one recipient receives the correct letter?
- Letters: 1, 2, 3. Recipients: A, B, C.
- Total number of outcomes: 3x2=6 (Are they equally likely?)
- Outcomes where only one person gets the right letter: (A-1, B-3, C-2), (A-3, B-2, C-1), (A-2, B-1, C-3)
- Outcomes where only two people get the right letter: 0
- Outcomes where everyone get the right letter: 1.
- Total probability = 4/6 = 2/3

General Problem-solving Strategies

- First, identify Ω, the set of all possible outcomes.
- Second, identify the event.
  - The set of outcomes where one or more people get their letter
- Third, assign the probability to the event
  - Using the probabilities of atomic events
  - [Optionally] using the assumption that all atomic events are equally likely.

Independence and Bayes' Rule

Conditional Probability

- Event A = {3}
- Event B = {3, 4, 6}
- What is the probability of A if we know B has happened?
- Since we know B has happened, the universe of outcomes is now {3, 4, 6} and they are equally probable.
- The probability of A is 1/3
- We write $P(A|B) = \frac{1}{3}$
- What is $P(B|A)$?

Conditional Probability

- Event A = {3}
- Event B = {3, 4, 6}
- For any event A and B,

$$\mathrm{P}(A \wedge B) = \mathrm{P}(A \cap B) = \mathrm{P}(\mathrm{A}, \mathrm{B}) = P(B)P(A|B) = \mathrm{P}(\mathrm{A})\mathrm{P}(\mathrm{B}|\mathrm{A})$$

- Verify for A = {3}, B = {3, 4, 6}
- P(AB) =
- P(B) P(A|B) =

Conditional Probability and Independence

- For any events A and B,

$$\mathrm{P(A, B)} = P(B)P(A|B) = \mathrm{P(A)P(B|A)}$$

- If A and B are **independent**, one event has no effects on the probability of the other.

$$P(\mathrm{A|B}) = P(A), \qquad \mathrm{P(A, B)} = P(B)P(A)$$

- Assuming a fair die, if the first roll gets 6, what is the probability of the second roll getting 6?

Conditional Probability and Independence

- For any events A and B,

$$\mathrm{P(A, B)} = P(B)P(A|B) = \mathrm{P(A)P(B|A)}$$

- If A and B are **independent**, one event has no effects on the probability of the other.

$$P(\mathrm{A|B}) = P(A), \qquad \mathrm{P(A, B)} = P(B)P(A)$$

- What is the probability of getting a double 6?

Conditional Probability and Independence

- For any events A and B,

$$\mathrm{P(A, B)} = P(B)P(A|B) = \mathrm{P(A)P(B|A)}$$

- If A and B are **independent**, one event has no effects on the probability of the other.

$$P(\mathrm{A|B}) = P(A), \qquad \mathrm{P(A, B)} = P(B)P(A)$$

- What is the probability of getting the sequence {5, 4, 6, 1, 6}?

Conditional Independence

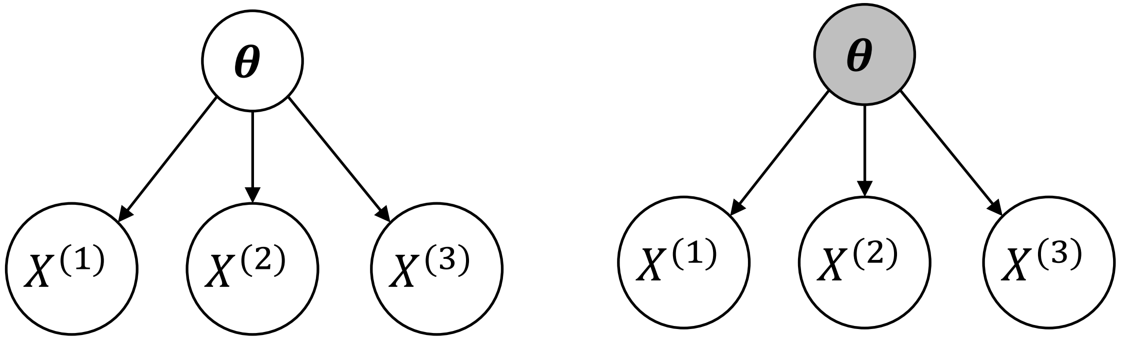
- **Assuming a fair die**, if the first roll gets 6, what is the probability of the second roll getting 6?
- **Without assuming a fair die**, if the first roll gets 6, what is the probability of the second roll getting 6?
- Answer: > 1/6
- Because the first roll indicates the die may be slightly biased toward 6.

Conditional Independence

- We write the probability associated with a die as a vector $\boldsymbol{\theta} = (\theta_1, \dots, \theta_6)$
- We write the result of the first and second rolls as $X^{(1)}$ and $X^{(2)}$
- $X^{(1)}$ and $X^{(2)}$ may be from 1 to 6
- $P\left(X^{(1)} \middle| \boldsymbol{\theta}\right) = (p_1, \dots, p_6)$
- Are $P(X^{(1)}|\boldsymbol{\theta})$ and $P(X^{(2)}|\boldsymbol{\theta})$ independent?
  - In other words, are $P(X^{(1)} = a|\boldsymbol{\theta})$ and $P(X^{(2)} = b|\boldsymbol{\theta})$ independent?
- Are $P(X^{(1)})$ and $P(X^{(2)})$ independent?
  - In other words, are $P(X^{(1)} = a)$ and $P(X^{(2)} = b)$ independent?

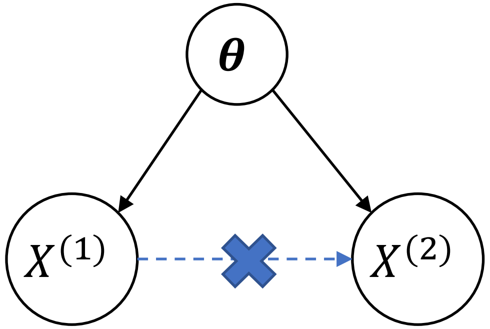
Conditional Independence

- **Conditional Independence**
- Given the probabilities of the 6 sides, different rolls are independent of each other.
- Without knowing the probabilities, the rolls are not independent.
- $P(X^{(1)}|\boldsymbol{\theta})$ and $P(X^{(2)}|\boldsymbol{\theta})$ are independent.
- $P(X^{(1)})$ and $P(X^{(2)})$ are not.



Probabilistic Graphic Models [Optional]

- Diagrams like these tell us how to write the joint distribution, which is the distribution that contains all variables.
- $P\left(X^{(1)}, X^{(2)}, \boldsymbol{\theta}\right) = P\left(X^{(1)} \middle| \boldsymbol{\theta}\right) P\left(X^{(2)} \middle| X^{(1)}, \boldsymbol{\theta}\right) P(\boldsymbol{\theta}) = P\left(X^{(1)} \middle| \boldsymbol{\theta}\right) P\left(X^{(2)} \middle| \boldsymbol{\theta}\right) P(\boldsymbol{\theta})$ (conditional independence between $X^{(1)}$ and $X^{(2)}$)
- The diagram encodes independence relations.



Notation

- For any events A and B,

$$\mathrm{P(A, B)} = P(B)P(A|B) = \mathrm{P(A)P(B|A)}$$

- For events A, B, and C

$$\mathrm{P(A, B, C)} = P(A|B, C)P(B|C)P(C)$$

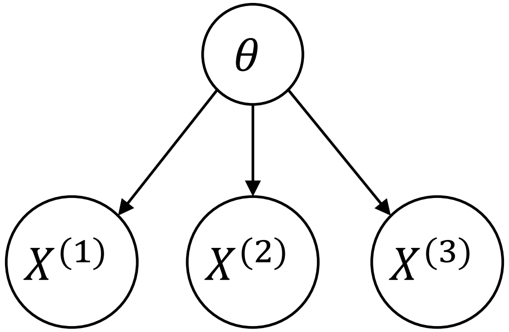
- Assuming A and B are independent given C

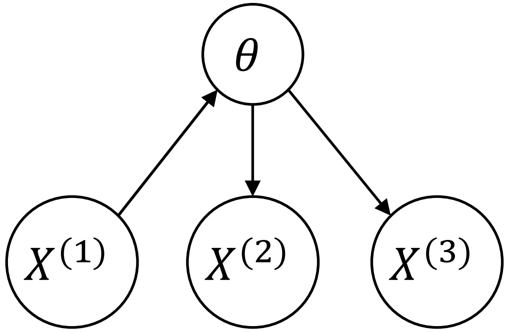
$$\mathrm{P(A, B, C)} = P(A|C)P(B|C)P(C)$$

- Note $P(C)$ only appears once in the above!

Directed Probability Graphs

- Graphs like these indicate the way we write the conditional distributions.
- The final goal is to write the full joint distribution $P(X^{(1)}, X^{(2)}, X^{(3)}, \theta)$
- The graph indicates that we factorize the joint distribution as
- $P\left(X^{(1)}, X^{(2)}, X^{(3)}, \theta\right) = P(X^{(1)}|\theta)P(X^{(2)}|\theta)P\left(X^{(3)}\middle|\theta\right)P(\theta)$
- We can use other factorizations but they may not be easy to specify.
- $P\left(\theta\middle|X^{(1)}\right) = ?$



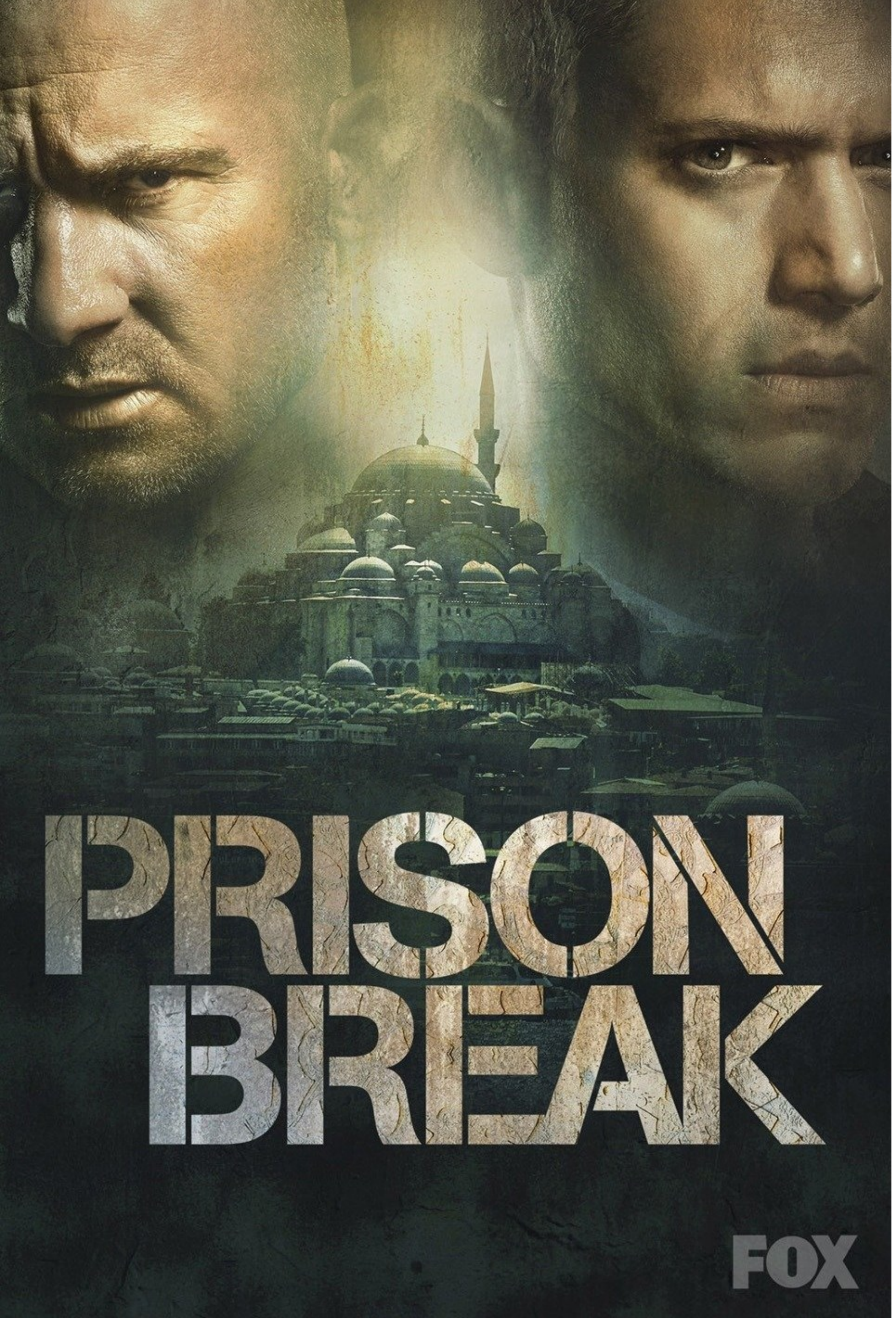


Bayes' Theorem

- A.k.a, Bayes' law or Bayes' rule

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

- How is this derived?



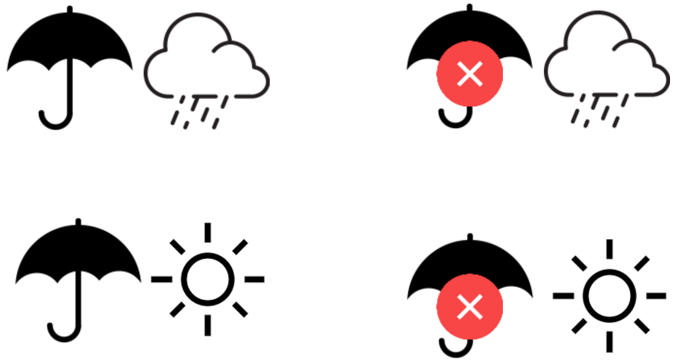
Bayes' Theorem for Statistical Inference

- You are locked in a prison cell without windows. You can observe a prison guard who sometimes carries an umbrella. You know that, if it rains outside, the probability of the guard carrying an umbrella is 0.8. If it does not rain, the probability of the guard carrying the umbrella is 0.1. The probability of raining in any given day is 0.5 in your climate.
- What is the probability of raining, if you observe the guard with an umbrella?

Bayes' Theorem

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

Sample Space



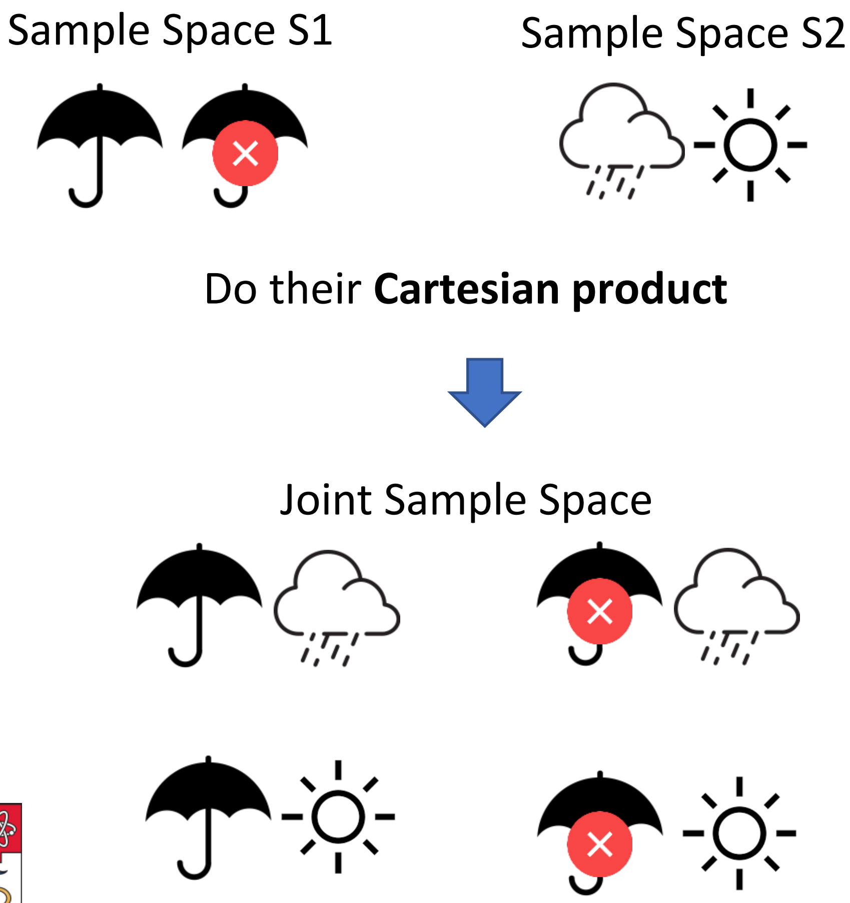
What is the probability of raining, if you observe the guard with an umbrella?

Conditional Probability Table

If ↓	Carry Umbrella	Not Carry Umbrella
Rain	0.8	1-0.8=0.2
Not Rain	0.1	0.9

	Rain	Not Rain
	0.5	0.5

Marginalization



Sample Space S1

Sample Space S2

Do their **Cartesian product**

Joint Sample Space

- P(☂) = P(☂ 🌧) + P(☂ ☀)
- This is known as marginalization.
- We eliminate the weather variable because we have considered all of its possibilities.
- That is, we marginalized the weather variable.

Bayes' Theorem

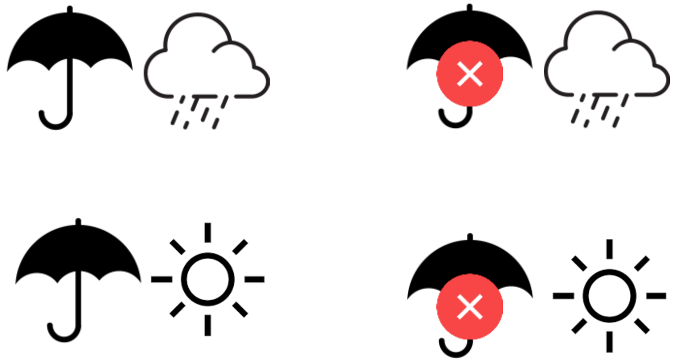
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Conditional Probability Table

If ↓	Carry Umbrella	Not Carry Umbrella
Rain	0.8	1-0.8=0.2
Not Rain	0.1	0.9

	Rain	Not Rain
	0.5	0.5

Sample Space



- $P(\text{Rain} \mid \text{Umbrella}) = \frac{P(\text{Umbrella}|\text{Rain})P(\text{Rain})}{P(\text{Umbrella})}$
- $P(\text{Umbrella}) = P(\text{Umbrella and Rain}) + P(\text{Umbrella and} \neg \text{ Rain})$
- $P(\text{Umbrella}) = \text{P}(\text{Umbrella} \mid \text{Rain})P(\text{Rain}) + \text{P}(\text{Umbrella} \mid \neg \text{ Rain})P(\neg\text{Rain})$
- $P(\text{Umbrella}) = 0.45$
- $P(\text{Umbrella}|\text{Rain})P(\text{Rain}) = 0.4$
- Therefore, $P(\text{Rain} \mid \text{Umbrella}) = 0.4/0.45 = 0.89$

Bayes' Theorem

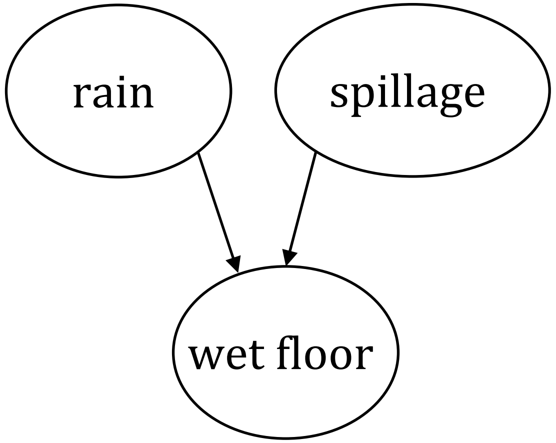
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Bayes' theorem is important in machine learning because we often do the following.

1. We design a model which gives us the probability P(data | model)
2. We observe some data.
3. We want to find out P(model | data). That is, what are some probable models of the world given the observations.

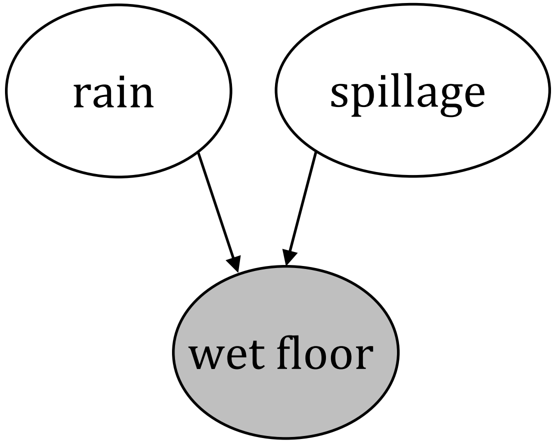
Independence: Example

- Raining causes the floor to be wet (with high probability).
- Someone spilling a soft drink on the floor causes it to be wet (with high probability).
- Are raining and spilling a drink independent?



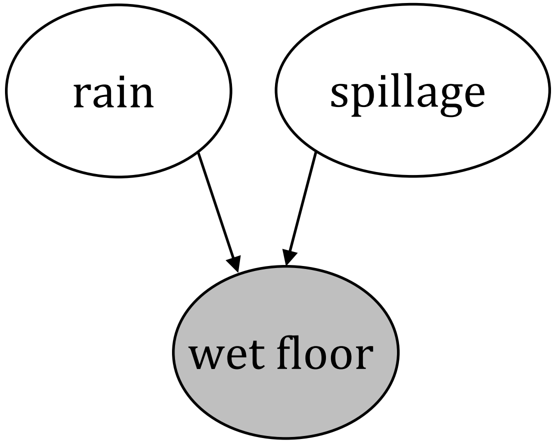
Independence: Example

- Raining causes the floor to be wet (with high probability).
- Someone spilling a soft drink on the floor causes it to be wet (with high probability).
- **If we know the floor is wet,** are raining and spilling a drink independent?

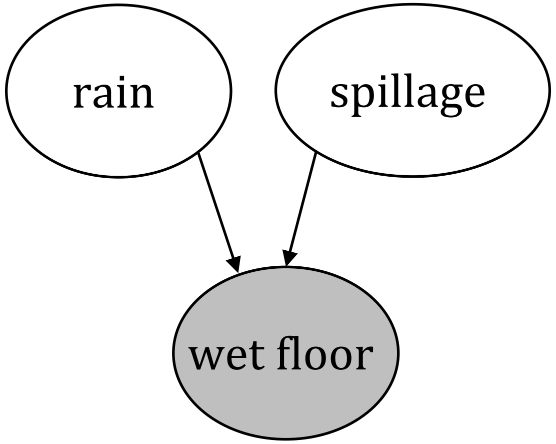


Independence: Example

- Raining causes the floor to be wet (with high probability).
- Someone spilling a soft drink on the floor causes it to be wet (with high probability).
- **If we know the floor is wet,** are raining and spilling a drink independent?
- They are no longer independent!
- Because the two reasons "compete" to explain the observed wet floor. If one of them is not true, the other is more likely to be true.



Independence: Example



Conditional Probability Table

If ↓	Wet Floor	Dry Floor
No Rain, No Spillage	0.05	0.95
No Rain, Spillage	0.6	0.4
Rain, No Spillage	0.7	0.3
Rain, Spillage	0.9	0.1

Synergy between the two

Prior Probabilities

Rain	No Rain
0.4	0.6

Spillage	No Spillage
0.2	0.8

$$P(\mathrm{R}|\mathrm{W},\neg\mathrm{S}) = \frac{\overset{\text{Joint distribution}}{P(\mathrm{R},\mathrm{W},\neg S)}}{\underbrace{P(\mathrm{R},\mathrm{W},\neg S) + P(\neg\mathrm{R},\mathrm{W},\neg S)}_{\text{marginalization}}}$$

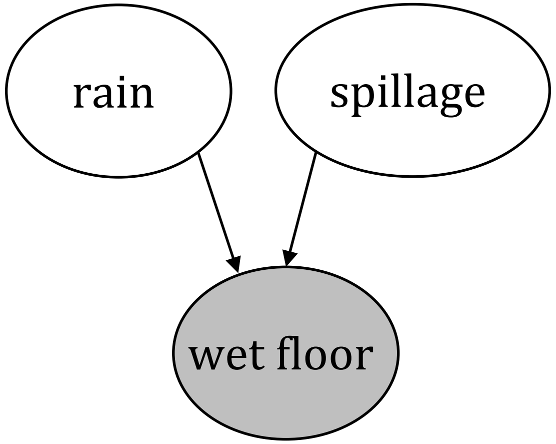
$$= \frac{0.4 \times 0.8 \times 0.7}{0.4 \times 0.8 \times 0.7 + 0.6 \times 0.8 \times 0.05}$$

$$= 0.9$$

$$P(\mathrm{R}|\mathrm{W}) = \frac{P(\mathrm{R},\mathrm{W})}{P(\mathrm{W})}$$

$$= \frac{P(\mathrm{R},\mathrm{W},\neg S) + P(\mathrm{R},\mathrm{W},S)}{P(\mathrm{R},\mathrm{W},\neg S) + P(\mathrm{R},\mathrm{W},S) + P(\neg\mathrm{R},\mathrm{W},\neg S) + P(\neg\mathrm{R},\mathrm{W},S)} = 0.755$$

Independence: Example



Conditional Probability Table

If ↓	Wet Floor	Dry Floor
No Rain, No Spillage	0.05	0.95
No Rain, Spillage	0.6	0.4
Rain, No Spillage	0.7	0.3
Rain, Spillage	0.9	0.1

Synergy between the two

Prior Probabilities

Rain	No Rain
0.4	0.6

Spillage	No Spillage
0.2	0.8

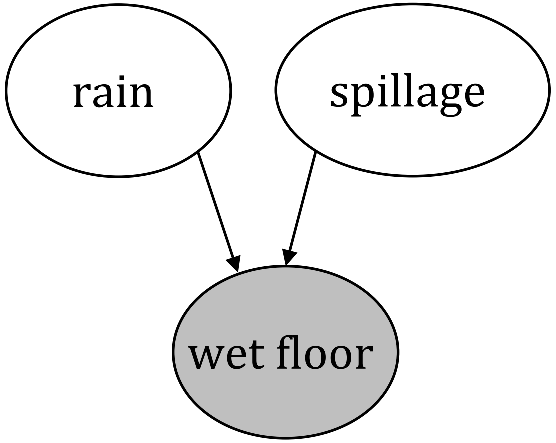
$$P(\mathrm{R}|\mathrm{W}, \neg\mathrm{S}) = 0.9$$

$$P(\mathrm{R}|\mathrm{W}) = 0.755$$

The two are not equal, so the two events are not independent when wet floor is given.

Independence: Example

- **If we know the floor is wet,** raining and spilling a drink are no longer independent!
- Any V-shape in directed graphs has this property.
- Often referred to as “explain away”.



Advanced Topics That are not Covered

- **Probabilistic Graphic Models**
    - Directed graphs
    - Undirected graphs
    - Factor graphs
- The graph representations encode independence relations between variables
- Central Question: How to find the distribution of unknown variables given the observed or known variables?
- These models were very popular before the era of deep learning

Random Variables and Their Distributions

Random Variable

- **Sample space** Ω: The set of possible outcomes ω of a random experiment
- A **random variable** is a real-valued function defined on a sample space.
- For the sample space {1, 2, 3, 4, 5, 6}, we may define R.V. X as
  - For example, 1 -> 1.5, 2 -> 100, 3 -> -0.32, 4 -> -3, 5 -> 5, 6 -> -6
  - We usually use something more intuitive. ;)
- A RV takes on some value depending on the outcome of the random experiment.
- Any real-valued function of an RV is also an RV

Random Variable: Example

- Flip a coin 3 times. Possible outcomes: {HHH, HHT, HTH, HTT, TTH, TTT}.
- Random variable: The number of heads in the outcome.
- It falls in the set {0, 1, 2, 3}.

Random Variable

- Sometimes we just treat some value as a random variable without specifying the "experiment" because that allows us to use the statistical tools.
  - The height or weight of a random Singaporean.
  - The position of a sub-atomic particle.
  - The stock price at 12.30 tomorrow.
  - Whether an image is a dog (X=1) or a cat (X=0).

Random Variable

- Functions of random variables are random variables. That is, they are composite functions of "random experiments".
- R.V. X can take on values {0, 1, 2, 3}
- Y=X+1 is another random variable that can take on values {1, 2, 3, 4}
- The average height of a sample group of Singaporeans.
  - Assume you randomly sample 10 Singaporeans and compute their average height. It is a function of 10 RV, which are functions of their own sample spaces. Thus, the average is a function of their joint sample space.
  - If you repeat the same sampling procedure (i.e., the random experiment), you will get different values.

Probabilistic Distribution

- Discrete: random variables can be integer-valued
- Continuous: random variables can be real-valued
- Multivariate or multidimensional: random variables can be vectors or matrices.

Probabilistic Distribution

- Discrete: random variables can be integer-valued (assuming from 1 to N)
- Technically, finite and countably infinite. In most cases, we deal with a finite range.
- Probability of each possible value ≥ 0

$$P(X = i) \geq 0$$

Probability mass function

- The sum of probabilities must be equal to 1.

$$\sum_{i=1}^{N} P(X = i) = 1$$

- The expectation / mean of the distribution

$$\mu = \sum_{i=1}^{N} i \, P(X = i)$$

- The variance of the distribution

$$\sigma = \sum_{i=1}^{N} (i - \mu)^2 P(X = i)$$

Bernoulli Distribution

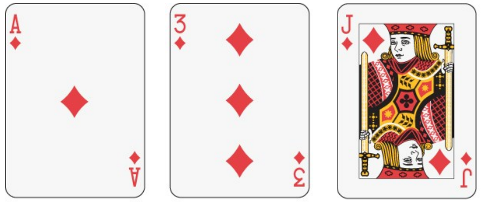
- The experiment has two outcomes, like flipping a coin, and is performed only once.
- RV X can take on two values, a and b (we use 1 and 0).
- One parameter: $1 \geq \theta \geq 0$
- $P(X = 1) = \theta$
- $P(X = 0) \ = \ 1 - \theta$
- $P(X = c) = 0, \forall c \neq 1 \text{ or } 0$
- Mean: θ
- Variance: $\theta(1 - \theta)$

Binomial Distribution

- Repeating the Bernoulli experiment for K times.
- RV X is the number of times we get 1 (or 0) in all trials.
- Say, flipping a coin K times and recording the total number of heads.
- Only one parameter: $1 \geq \theta \geq 0$
- $P(X = i) = \binom{K}{i} \theta^i (1 - \theta)^{K-i}$
- Mean: $K\theta$
- Variance: $K\theta(1 - \theta)$

Categorical Distribution

- The experiment has N>2 possible outcomes. This is like rolling a 6-sided die or drawing a poker card for once.
- RV X can take on multiple values (e.g, 1 to N).
- Suppose we draw 1 card from the suit of diamond, including cards from A to K.
- If the probability of every outcome is equal, then every outcome has the probability of $\frac{1}{13}$



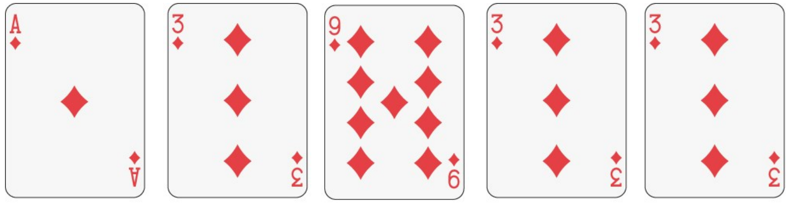
Categorical Distribution

- The experiment has N>2 possible outcomes. This is like rolling a 6-sided die or drawing a poker card for once.
- RV X can take on multiple values (e.g, 1 to N).
- **The probabilities of different outcomes need not be equal.**
- N-1 free parameters: θ_1 to θ_{n-1}. θ_n is determined once the first N-1 are determined.
- $\forall i, \theta_i \geq 0, \sum_{i=1}^{N} \theta_i = 1$
- $P(X = i) = \theta_i$

Multinomial Distribution

- The experiment has N possible outcomes. It is repeated K times.
- RV X_1 … X_N denote the number of times we observe the n^{th} outcome. $\sum_{i=1}^{N} X_i = K$
- Suppose we sequentially draw 5 cards from the suit of diamond, including cards from A to K. After drawing a card, we put it back in the pile.

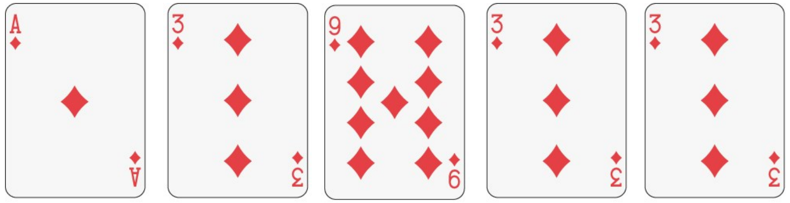
The sequence of card drawn:



Multinomial Distribution

- RV $X_1 \ldots X_N$ denote the number of times we observe the n^{th} outcome. $\sum_{i=1}^{N} X_i = K$

The sequence of card drawn:



$$X_1 = 1, X_2 = 0, X_3 = 3, \ldots, X_9 = 1, X_{10} = 0, \ldots, X_{13} = 0$$

If every card is equally likely, the probability of observing the above sequence (or any sequence) is $\left(\frac{1}{13}\right)^5$

Discrete Probabilistic Distribution: Multinomial

- Multinomial Distribution. The experiment has N possible outcomes. It is repeated K times.
- RV $X = \langle X_1 \ldots X_N \rangle$ denote the number of times we observe the i^{th} outcome. $\sum_{i=1}^{N} X_i = K$
- **The probabilities of different outcomes need not be equal.**
- N-1 free parameters: θ_1 to θ_{n-1}. θ_n is determined once the first N-1 are determined.
- $\forall i, \theta_i \geq 0, \sum_{i=1}^{N} \theta_i = 1$
- For each possible value i, we observe it happening X_i times
- The total probability is

$$\frac{K!}{\prod_i^N (X_i!)} \prod_{i=1}^{N} \theta_i^{X_i}$$

Estimating a probability distribution

- We have a coin. We don’t know if it is fair.
- We toss it 100 times and get Head 56 times.
- What is our best estimate for the probability of getting Head?
- It depends on what “best” means.
- But without additional information, the most appropriate answer is 56/100 = 0.56
- <u>This is different from the question: How confidence are we about the claim that it is fair/not fair?</u>

Maximum Likelihood Estimation

- We have a probability distribution $P(X|\theta)$, which is about random variable X and parameterized by θ.
- We have K observations, $X^{(1)}, \dots, X^{(K)}$
- We want to find θ
- Claim: The best θ is the one that maximizes $P(X^{(1)}, \dots, X^{(K)}|\theta)$

$$\hat{\theta}_{\text{MLE}} = \underset{\theta}{\operatorname{argmax}} P(X^{(1)}, \dots, X^{(K)}|\theta)$$

Maximum Likelihood Estimation: Example

- Binomial distribution $P(X;\theta)$
- Observation: 1 = 56 times, 0 = 44 times.

$$P\left(X^{(1)}, \dots, X^{(K)} | \theta\right) = \binom{100}{56}\theta^{56}(1-\theta)^{44}$$

- How to find the maximum over p?

$$\hat{p}_{\mathrm{MLE}} = \underset{\theta}{\operatorname{argmax}} \binom{100}{56}\theta^{56}(1-\theta)^{44}$$

$$= \underset{\theta}{\operatorname{argmax}} \log\left(\binom{100}{56}\theta^{56}(1-\theta)^{44}\right)$$

Logarithm is monotonic.

$$= \underset{\theta}{\operatorname{argmax}}\, 56\log(\theta) + 44\log(1-\theta)$$

The term $\binom{100}{56}$ has nothing to do with θ

Maximum Likelihood Estimation: Example

- Find the maximum of

$$56\log(\theta) + 44\log(1-\theta)$$

- Setting the derivative to zero

$$\frac{56}{\theta} - \frac{44}{1-\theta} = 0$$

$$56(1-\theta) - 44\theta = 0$$

$$56 - 56\theta - 44\theta = 0$$

$$\theta = \frac{56}{100}$$

You can verify that this is a maximum, not a minimum, by plugging in 0.57 and 0.55.

The MLE agrees with our intuition.

Continuous Random Variables

- Continuous RVs are more difficult to deal with because the number of possibilities are infinite and uncountable.
  - For example, RV X can take any value on the real line
- Cumulative probability function $F_X(a) = P(X \leq a)$
  - $F_X(\cdot)$ is non-decreasing, i.e., $F_X(a) \geq F_X(b)$ if $a \geq b$
  - $\lim_{x \to -\infty} F_X(x) = 0, \lim_{x \to \infty} F_X(x) = 1$
  - $F_X(\cdot)$ is right-continuous, i.e., $\lim_{x \to a^+} F_X(x) = F_X(a)$

Continuous Random Variables

- For continuous RV X

$$F_X(x) = \int_{-\infty}^{x} f_X(y)dy$$

Why do we write y here?

cumulative probability function

probability density function

- $f_X(x) \geq 0, \forall x$
- $\int_{-\infty}^{\infty} f_X(x)dx = 1$
- $\frac{d}{dx} F_X(x) = f_X(x)$

Fundamental Theorem of Calculus

Gaussian Distribution

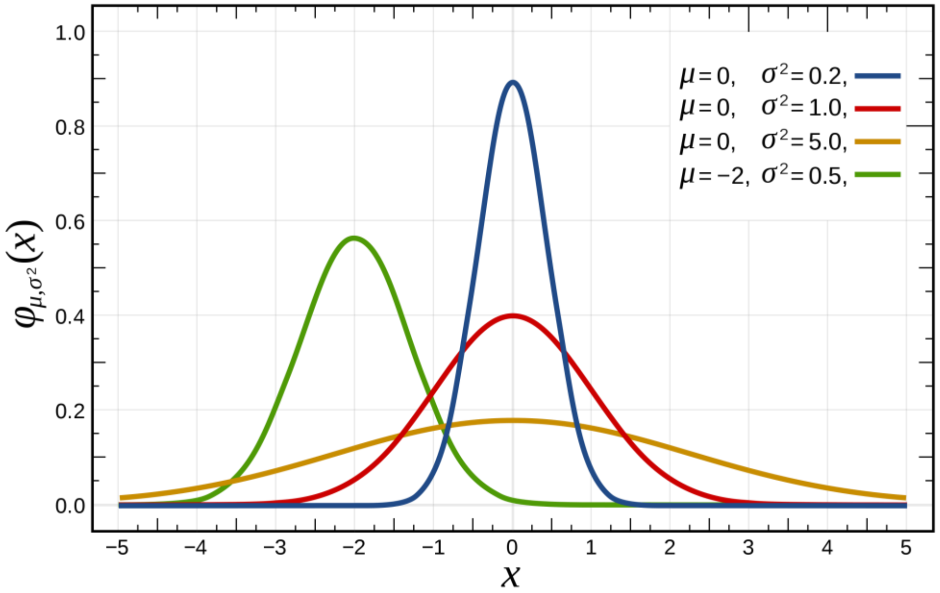
- A.k.a normal distribution, bell curve
- The most frequently used distribution
- Two parameters: mean μ and standard deviation σ

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

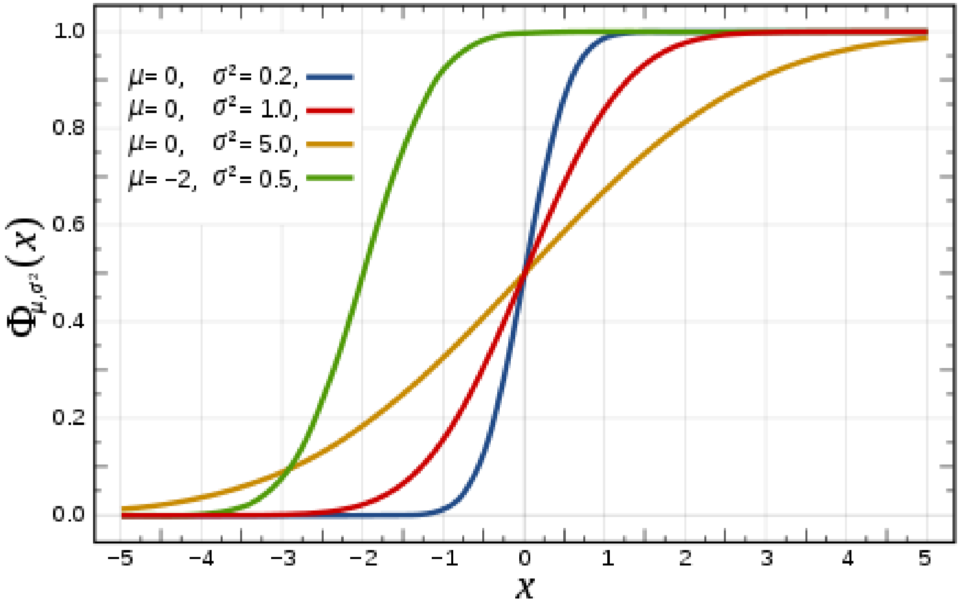
Probability density function

Gaussian Distribution

- A.k.a normal distribution, bell curve
- The most frequently used distribution



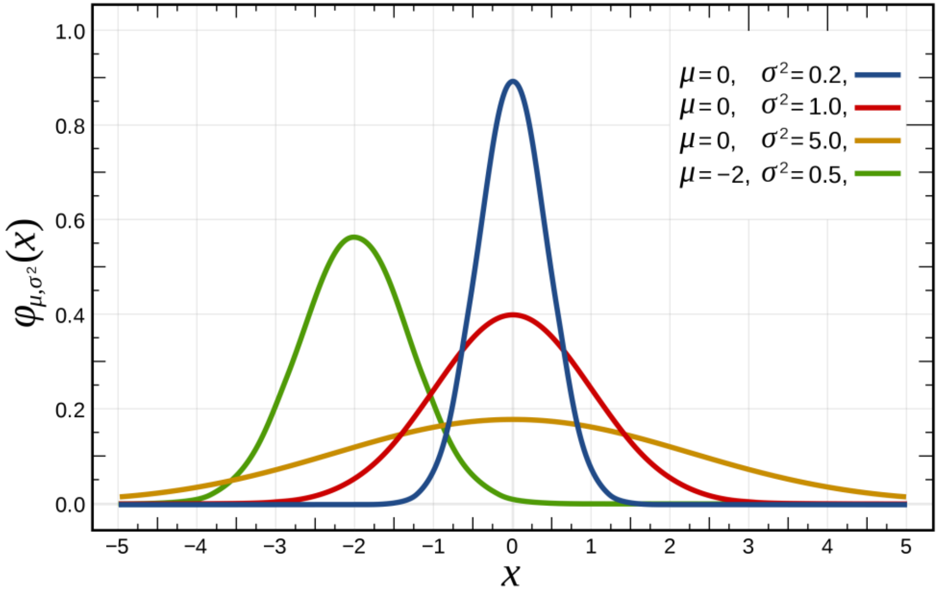
Probability density function



Cumulative probability function

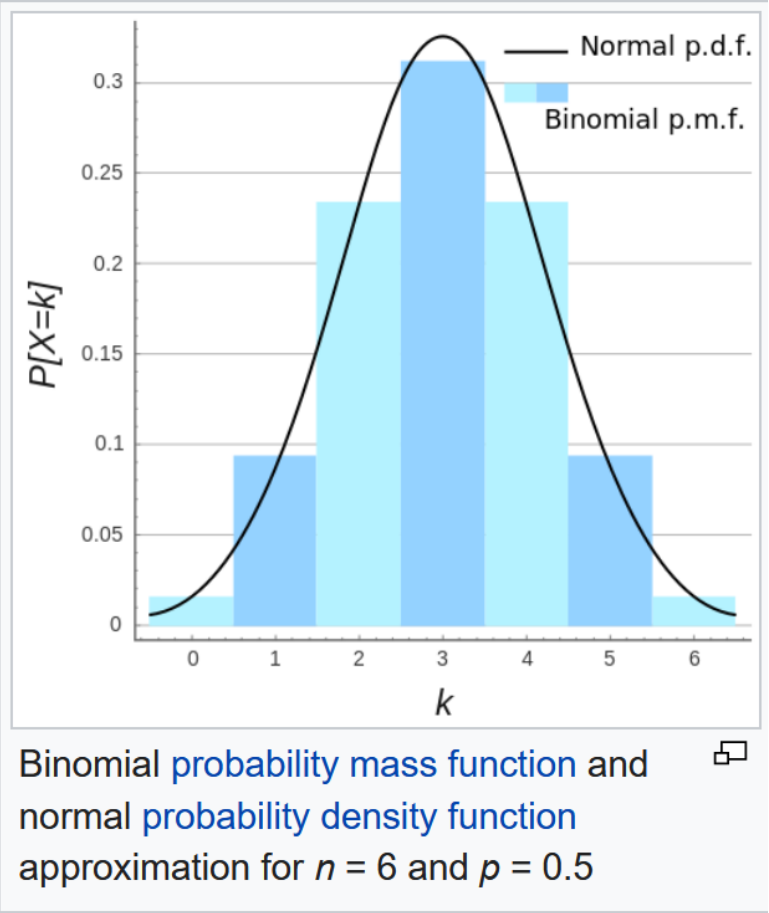
Gaussian Distribution

- The Gaussian distribution is widely used in statistics and can describe many different natural phenomena
  - Height, weight, measurement errors, IQ, etc.
- Due to the central limit theorem, the overall effect of many small additive factors tend to become normally distributed.



Gaussian Distribution

- When the number of trials, n, is large, the binomial distribution can be approximated by the Gaussian distribution.



Binomial probability mass function and normal probability density function approximation for $n = 6$ and $p = 0.5$

Maximum Likelihood Estimation

- We have N observations, $X^{(1)}, \dots, X^{(N)}$
- We want to find μ, σ
- Claim: The best μ, σ are the ones that maximize $P\left(X^{(1)}, \dots, X^{(N)} \middle| \mu, \sigma\right)$

$$\hat{\theta}_{\text{MLE}} = \underset{\theta}{\operatorname{argmax}}\, P\left(X^{(1)}, \dots, X^{(N)} \middle| \mu, \sigma\right)$$

$$= \underset{\theta}{\operatorname{argmax}} \prod_{i}^{N} P\left(X^{(i)} \middle| \mu, \sigma\right)$$

The observations are independent given μ, σ

$$= \underset{\theta}{\operatorname{argmax}} \sum_{i}^{N} \log P\left(X^{(i)} \middle| \mu, \sigma\right)$$

MLE: Mean μ

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

$$\hat{\mu}_{\text{MLE}} = \underset{\mu}{\operatorname{argmax}} \sum_{i=1}^{N} \log P\left(X^{(i)} \middle| \mu, \sigma \right)$$

$$= \underset{\mu}{\operatorname{argmax}} \sum_{i=1}^{N} \log \frac{1}{\sigma\sqrt{2\pi}} + \log \exp - \frac{\left(X^{(i)} - \mu\right)^2}{2\sigma^2}$$

$$= \underset{\mu}{\operatorname{argmax}} \sum_{i=1}^{N} - \frac{\left(X^{(i)} - \mu\right)^2}{2\sigma^2}$$

Setting the derivative to zero

$$\frac{d}{d\mu} \sum_{i=1}^{N} - \frac{\left(X^{(i)} - \mu\right)^2}{2\sigma^2} = \sum_{i=1}^{N} \frac{\left(X^{(i)} - \mu\right)}{2\sigma^2} = 0$$

$$\sum_{i=1}^{N} X^{(i)} = N\mu$$

$$\hat{\mu}_{\text{MLE}} = \frac{1}{N} \sum_{i=1}^{N} X^{(i)}$$

The MLE agrees with our intuition.

MLE: Standard Deviation σ

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

$$\hat{\sigma}_{\text{MLE}} = \underset{\sigma}{\operatorname{argmax}} \sum_{i=1}^{N} \log P\left(X^{(i)} \middle| \mu, \sigma\right)$$

$$= \underset{\sigma}{\operatorname{argmax}} \sum_{i=1}^{N} \log \frac{1}{\sigma\sqrt{2\pi}} + \log \exp - \frac{\left(X^{(i)} - \mu\right)^2}{2\sigma^2}$$

$$= \underset{\sigma}{\operatorname{argmax}} -N \log \sigma - N \log\sqrt{2\pi} + \sum_{i=1}^{N} - \frac{\left(X^{(i)} - \mu\right)^2}{2\sigma^2}$$

Setting the derivative to zero

$$\frac{d}{d\sigma}\left(-N \log \sigma - \sum_{i=1}^{N} \frac{\left(X^{(i)} - \mu\right)^2}{2\sigma^2}\right) = 0$$

$$-\frac{N}{\sigma} + \frac{\sum\left(X^{(i)} - \mu\right)^2}{\sigma^3} = 0$$

$$\hat{\sigma}_{\text{MLE}} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (X^{(i)} - \mu)^2}$$

The MLE agrees with our intuition.

Expectation of a Function (Continuous RV)

- A.k.a expected value
- Expectation of a random variable, also known as the mean

$$E_{P(x)}[x] = \int x\, P(x) dx$$

$P(x)$ is the PDF of the distribution.

- Expectation of x^2

$$E_{P(x)}[x^2] = \int x^2\, P(x) dx$$

- Expectation of $f(x)$

$$E_{P(x)}[f(x)] = \int f(x) P(x) dx$$

Expectation of a Function (Discrete RV)

- A.k.a expected value
- Expectation of a random variable, also known as the mean

$$E_{P(x)}[x] = \sum_i x_i \, P(X = x_i)$$

- Expectation of x^2

$$E_{P(x)}[x^2] = \sum_i x_i^2 \, P(X = x_i)$$

- Expectation of $f(x)$

$$E_{P(x)}[f(x)] = \sum_i f(x_i) \, P(X = x_i)$$

Expectation of a Function

- A.k.a expected value
- Expectation of a random variable, also known as the mean

$$E_{P(x)}[x] = \int x\, P(x) dx$$

$P(x)$ is the PDF of the distribution.

- Expectation of x^2

$$E_{P(x)}[x^2] = \int x^2\, P(x) dx$$

- Variance of x can be written as an expectation or a difference between an expectation and the square of another expectation

$$Var(x) = E[(x - E[x])^2] = E[x^2] - (E[x])^2$$

Estimators

- An estimator is a method for calculating an estimate of a RV based on observed data.
    - Estimators for μ and σ of a Gaussian distribution
- An estimator of an RV is an RV!
- The estimator is a function of repeated outcomes of a random experiment. For example, rolling a die for 100 times. You can also think of it as the outcome of a large random experiment.
- How do we know if an estimator of an RV is any good?

Estimators

- We randomly sample 100 Singaporeans and get their average height, $\hat{\mu}^{(1)}$
- Repeat this procedure, we get $\hat{\mu}^{(2)}, \hat{\mu}^{(3)}, \hat{\mu}^{(4)}, \ldots$
- These $\hat{\mu}$ form a distribution!
- We will examine the mean and variance of this distribution, called the sampling distribution.

Bias and Variance

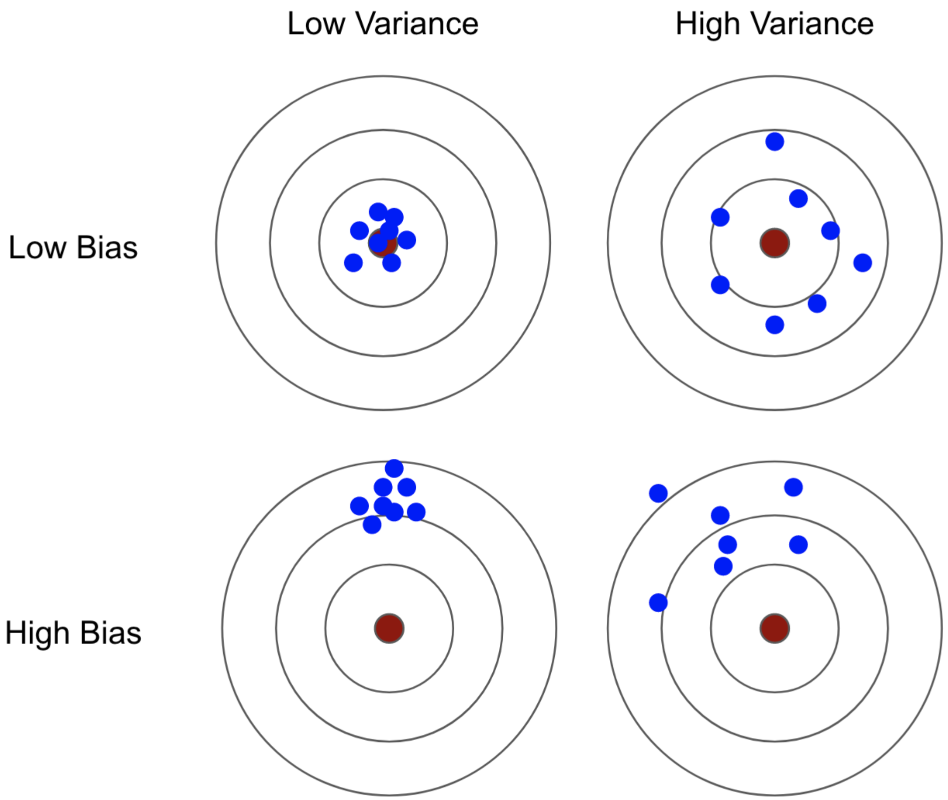
- How do we know if an estimator of an RV is any good?
- Bias: the difference between this estimator's expected value and the true value of the parameter being estimated.

$$E(\hat{\theta}) - \theta$$

- Variance: The scatter of the estimator around its expected value

$$E\left[\left(\hat{\theta}^{(i)} - E(\hat{\theta})\right)^2\right]$$

- Bias-variance trade-off: Often we can decrease bias at the cost of variance and vice versa.



Is $\hat{\mu}_{\mathrm{MLE}}$ Biased?

- $\hat{\mu}_{\mathrm{MLE}} = \frac{1}{N}\sum_{i=1}^{N} X^{(i)}$
- $E[\hat{\mu}_{\mathrm{MLE}}] = E\left[\frac{1}{N}\sum_{i=1}^{N} X^{(i)}\right] = \frac{1}{N}\sum_{i=1}^{N} E[X^{(i)}] = \frac{1}{N}\sum_{i=1}^{N} \boxed{\mu}$

mean of the real distribution of X

- $E[\hat{\mu}_{\mathrm{MLE}}] = \mu$
- Therefore, it's unbiased.
- Interestingly, $\hat{\sigma}_{\mathrm{MLE}}$ is biased.
- The unbiased estimator $\hat{\sigma}_{\mathrm{UB}} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(X^{(i)} - \mu)^2}$

What is the variance of $\hat{\mu}_{\mathrm{MLE}}$?

$$Var(\hat{\mu}_{\mathrm{MLE}}) = Var\left(\frac{1}{N}\sum_{i=1}^{N} X^{(i)}\right) = \frac{1}{N^2} Var\left(\sum_{i=1}^{N} X^{(i)}\right)$$

$$= \frac{1}{N^2}\sum_{i=1}^{N} Var\left(X^{(i)}\right) + \boxed{\sum_{i=1}^{N}\sum_{j=1}^{N} Cov\left(X^{(i)}, X^{(j)}\right)}$$

Independent variables have zero covariance

$$= \frac{1}{N^2} N\sigma^2 = \frac{\sigma^2}{N}$$

The variance of $\hat{\mu}_{\mathrm{MLE}}$ is proportional to the variance of the distribution of X and inversely proportional to the number of data points N

Advanced Topics That are Not Covered

- Other Commonly Used Distributions
  - Poisson Distribution
  - Exponential Distribution
  - Beta Distribution
  - Dirichelet Distribution
  - Fat-tailed Distributions, and many more
- Mixture of Distributions
  - Gaussian Mixtures, often used as a clustering model
- Bayesian Estimates
  - Incorporate “prior” into the estimates
- Estimators and Variance reduction
  - Variance is sometimes a problem (for example, in stochastic gradient descent)
  - Ways to reduce the variance of an estimator