

# UVA CS 4774: Machine Learning

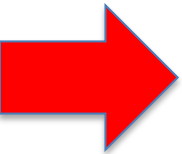
□

## Lecture: Probability Review

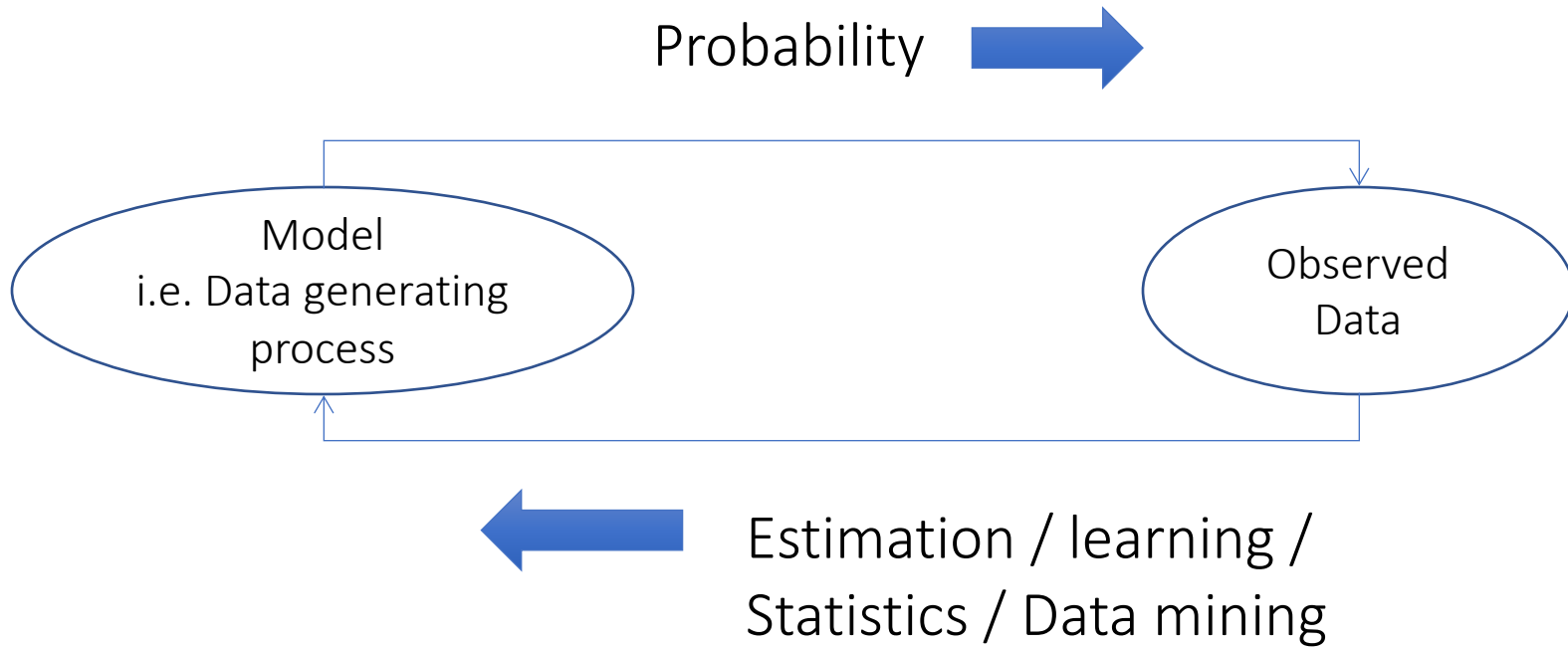
Dr. Yanjun Qi

University of Virginia  
Department of Computer Science

# Today : Probability Review

- 
- The big picture
  - Events and Event spaces
  - Random variables
  - Joint probability, Marginalization, conditioning, chain rule, Bayes Rule, law of total probability, etc.
  - Structural properties, e.g., Independence, conditional independence
  - Maximum Likelihood Estimation

# The Big Picture



# Probability

- Counting
- Basics of probability
- Conditional probability
- Random variables
- Discrete and continuous distributions
- Expectation and variance
- Tail bounds and central limit theorem
- .....

# Statistics

- Maximum likelihood estimation
- Bayesian estimation
- Hypothesis testing
- Linear regression
- [Machine learning]
- .....

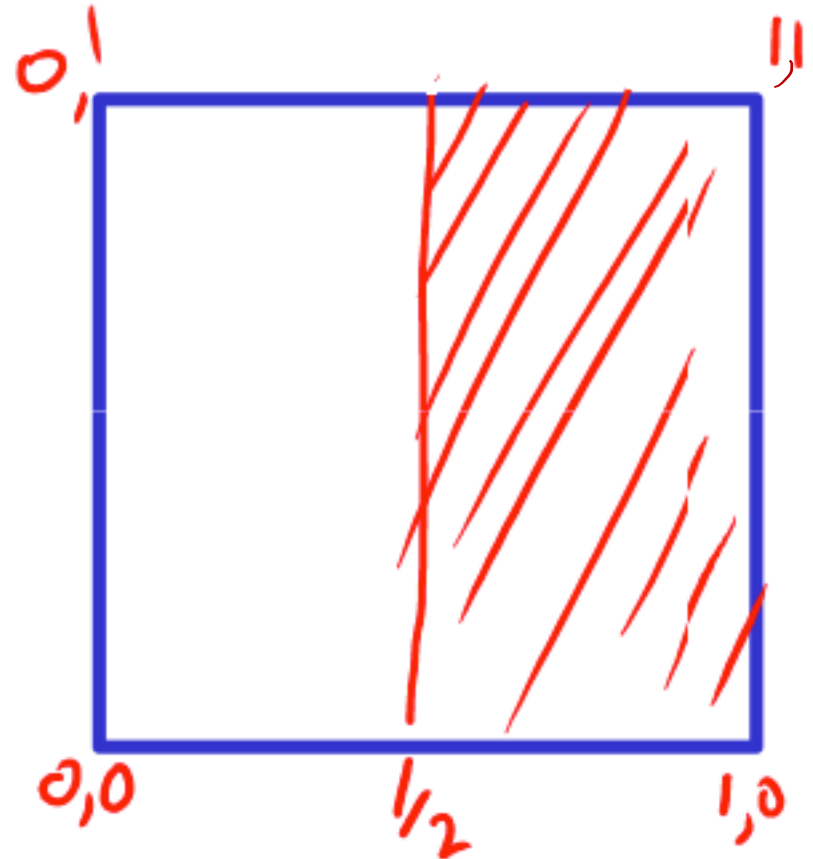
# Probability as frequency

- Consider the following questions:
  - 1. What is the probability that when I flip a coin it is “heads”?
  - 2. why ? We can count →  $\sim 1/2$
  - 3. What is the probability of Blue Ridge Mountains to have an erupting volcano in the near future ?  
→ could not count

Message: The frequentist view is very useful, but it seems that we can also use domain knowledge to come up with probabilities.

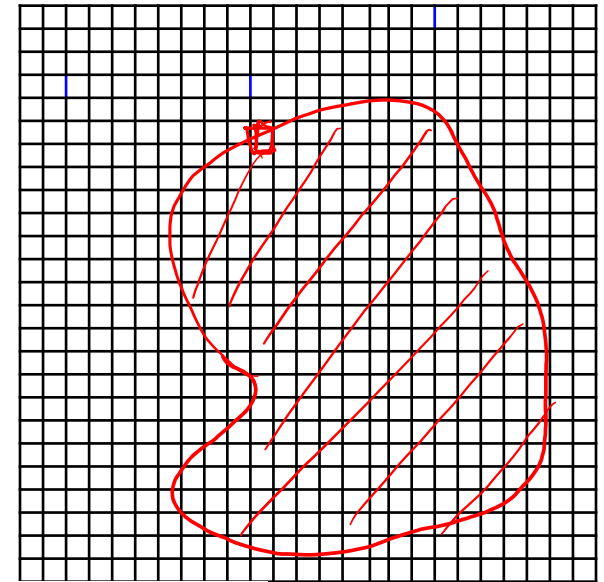
# Probability as a measure of uncertainty

- Imagine we are throwing darts at a wall of size  $1 \times 1$  and that all darts are guaranteed to fall within this  $1 \times 1$  wall.
- What is the probability that a dart will fall in the shaded area?



# Probability as a measure of uncertainty

- Probability is a **measure of certainty of an event taking place.**
- i.e. in the example, we were measuring the chances of hitting the shaded area.

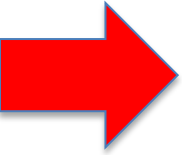


Its area is 1

$$prob = \frac{\# RedBoxes}{\# Boxes}$$



# Today : Probability Review

- 
- The big picture
  - Events and Event spaces
  - Random variables
  - Joint probability, Marginalization, conditioning, chain rule, Bayes Rule, law of total probability, etc.
  - Structural properties, e.g., Independence, conditional independence
  - Maximum Likelihood Estimation

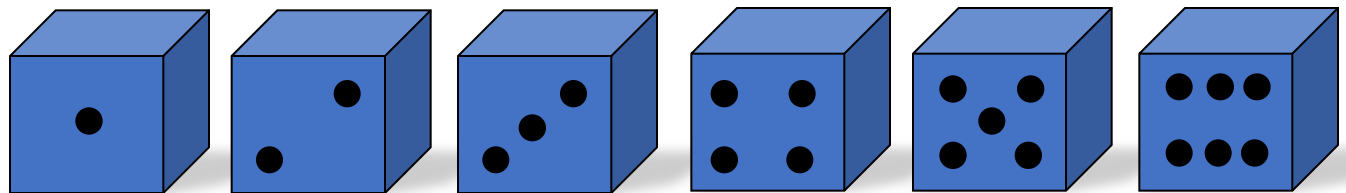
# Probability

**Probability** is the formal study of the laws of chance. Probability allows us to **manage uncertainty**.

The **sample space** is the set of all **outcomes**. For example, for a die we have 6 outcomes:

$$O_{\text{die}} = \{1, 2, 3, 4, 5, 6\}$$

O:



Elementary Event “Throw 2”

The elements of  $O$  are called elementary events.

# Probability

- *Probability allows us to measure many **events**.*
- *The **events** are subsets of the sample space  $\Omega$ . For example, for a die we may consider the following events: e.g.,*

$$\text{GREATER} = \{5, 6\}$$

$$\text{EVEN} = \{2, 4, 6\}$$

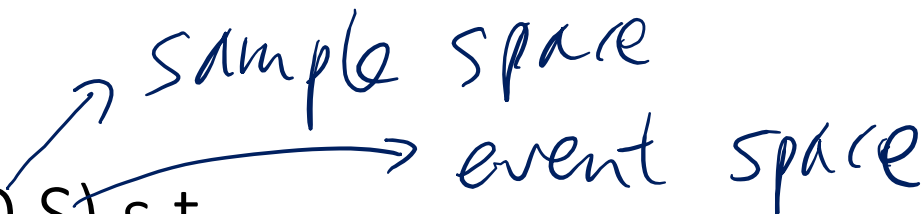
- *Assign probabilities to these events: e.g.,*

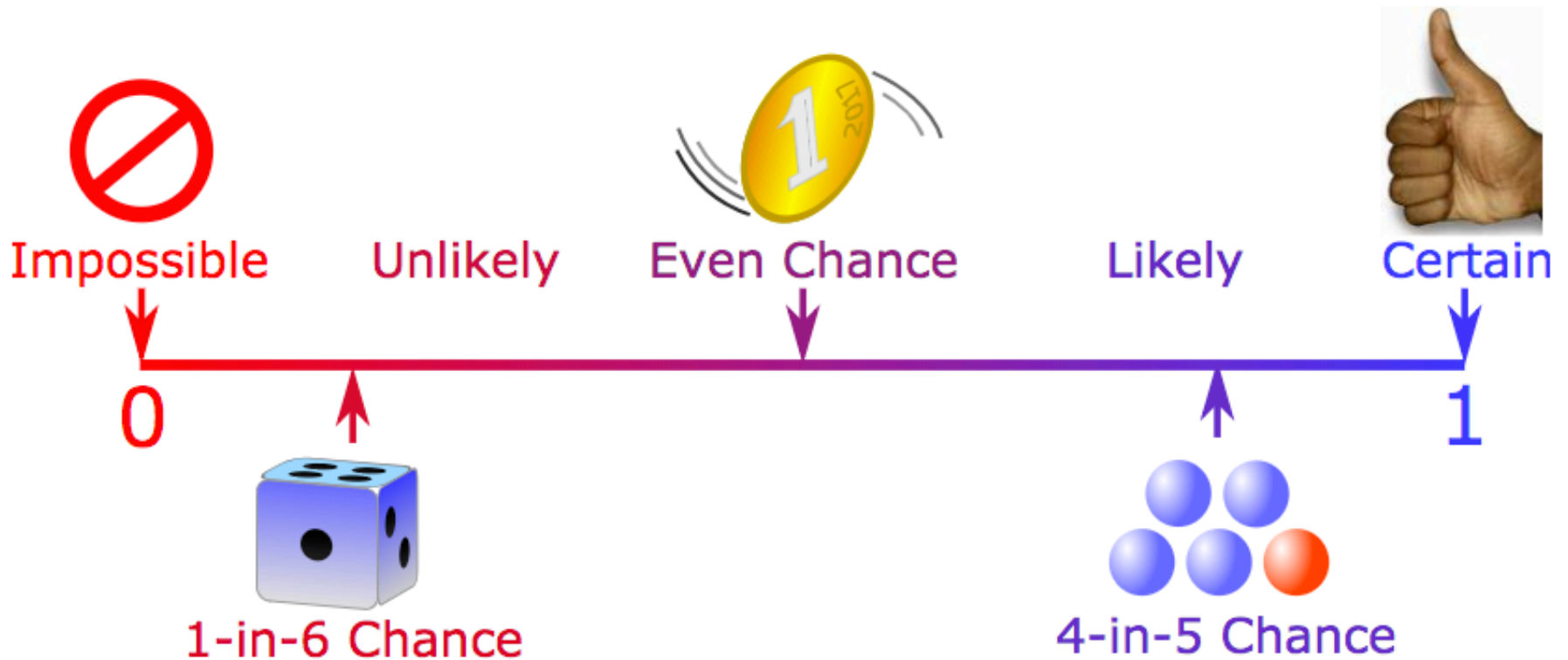
$$P(\text{EVEN}) = 1/2$$

# Sample space and Events

- $\Omega$  : **Sample Space**,
  - result of an experiment / set of all outcomes
  - If you toss a coin **twice**  $\Omega = \{HH, HT, TH, TT\}$
- **Event**: a subset of  $\Omega$ 
  - First toss is head =  $\{HH, HT\}$
- $\mathcal{S}$ : **event space, a set of events**:
  - Contains the empty event and  $\Omega$

# Axioms for Probability

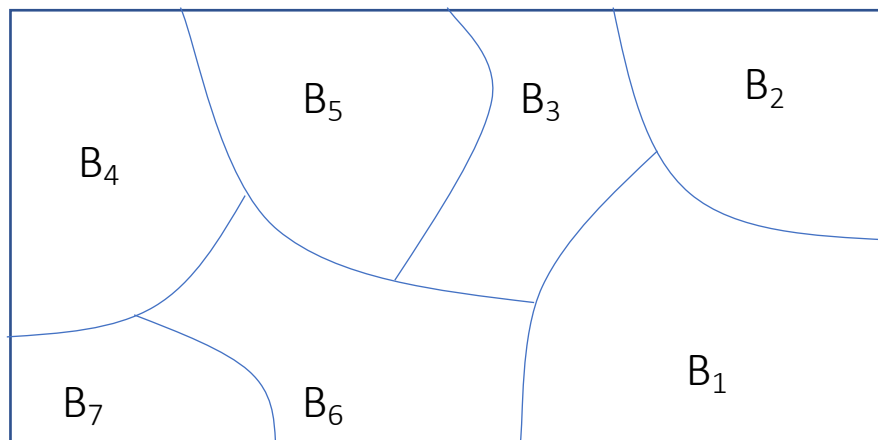
- Defined over  $(O, S)$  s.t. 
  - $1 \geq P(a) \geq 0$  for all  $\underbrace{a}$  in  $\underbrace{S}$
  - $P(O) = 1$
- If A, B are **disjoint**, then
  - $P(A \cup B) = p(A) + p(B)$



Probability is always between 0 and 1

# Axioms for Probability

$$\bullet P(\Omega) = \sum P(B_i) = 1$$

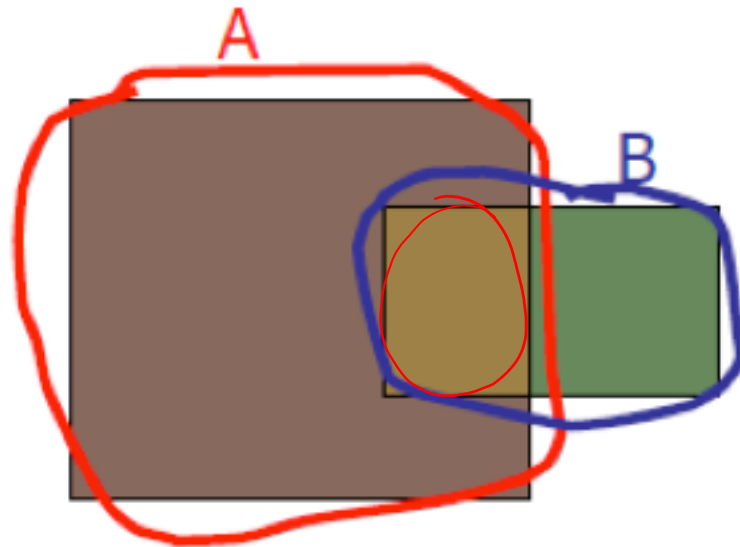


# OR operation for Probability

- We can deduce other axioms from the above ones
  - Ex:  $P(A \cup B)$  for **non-disjoint** events

$$P(\text{A or B}) = P(\text{A}) + P(\text{B}) - P(\text{A and B})$$

P( Union of A set and B set)



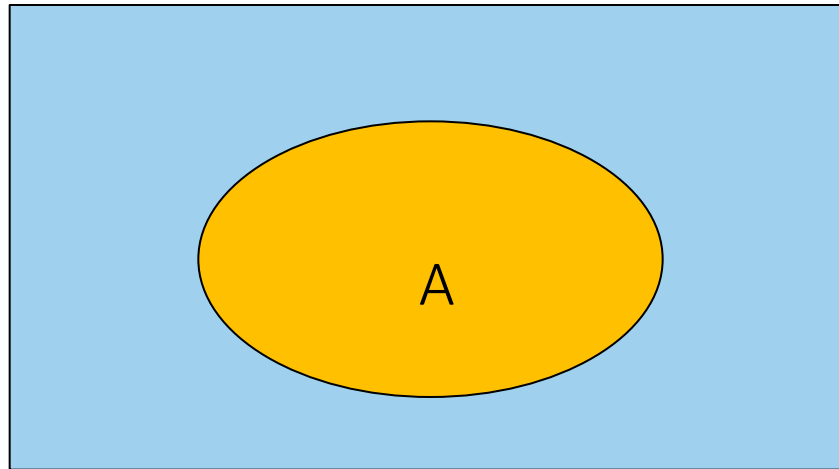


# NOT operation for Probability

- $0 \leq P(A) \leq 1$ ,
- $P(\text{A or B}) = P(\text{A}) + P(\text{B}) - P(\text{A and B})$

From these we can prove:

$$P(\text{not } A) = P(\sim A) = 1 - P(A)$$



# Law of Total Probability

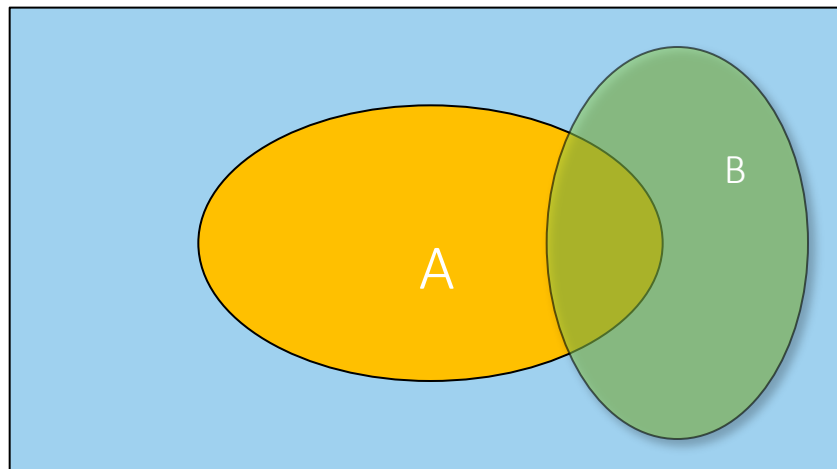
- $0 \leq P(A) \leq 1$ ,
- $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

From these we can prove:

$$P(A) = P(A \wedge B) + P(A \wedge \sim B)$$



P( Intersection of A and B)

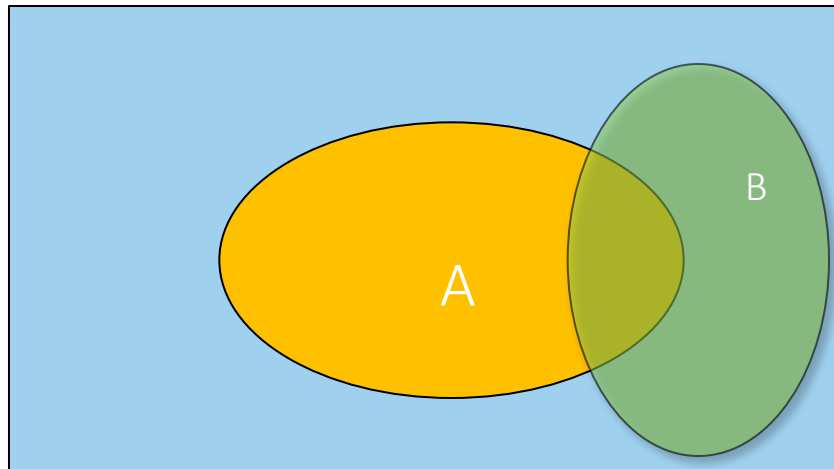


# Law of Total Probability

- $0 \leq P(A) \leq 1$ ,
- $P(\text{A or B}) = P(\text{A}) + P(\text{B}) - P(\text{A and B})$

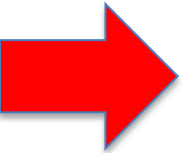
From these we can prove:

$$P(A) = P(A \cap B) + P(A \cap \sim B)$$



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

# Today : Probability Review

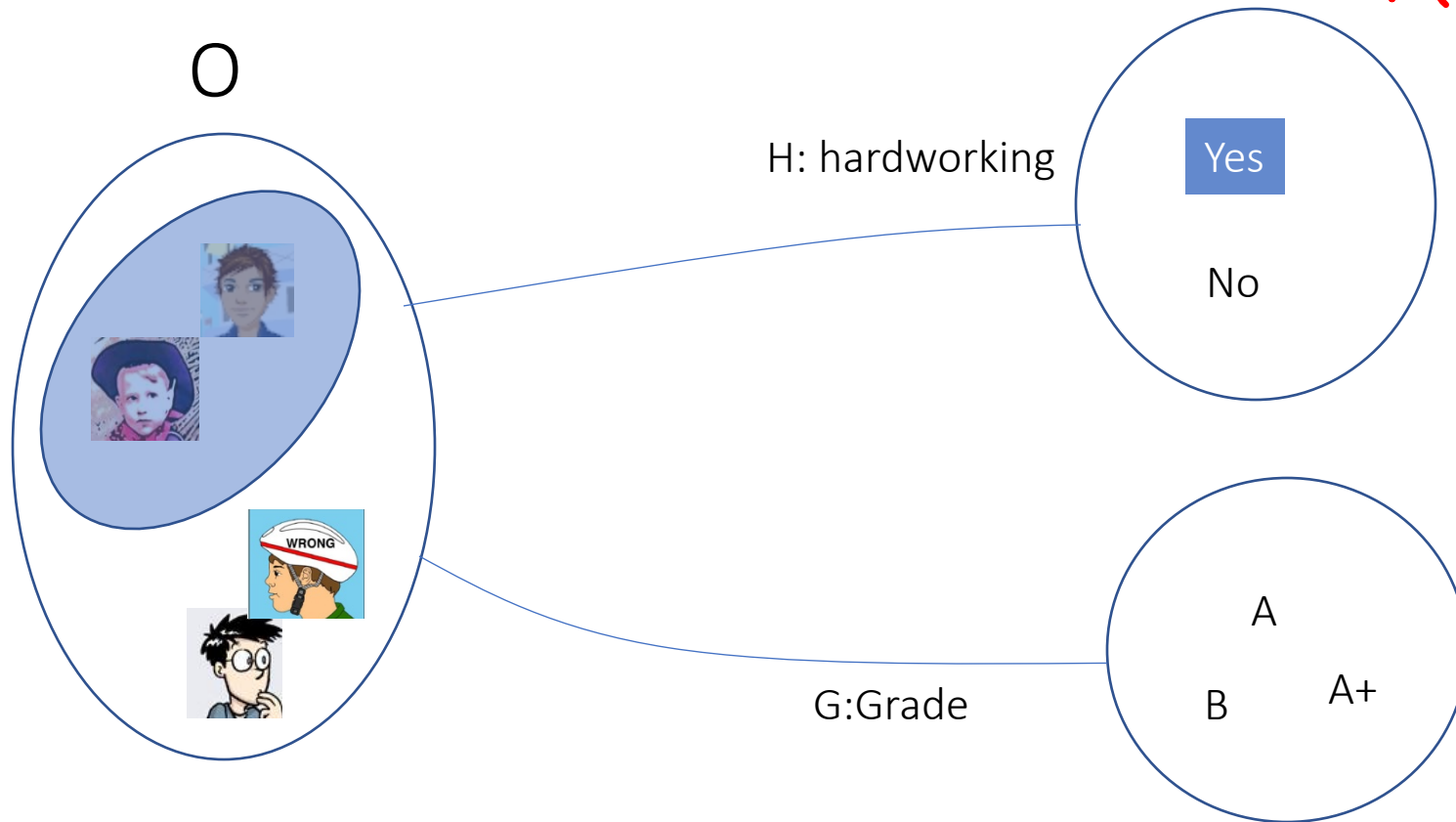
- The big picture
- Events and Event spaces
- • Random variables
- Joint probability, Marginalization, conditioning, chain rule, Bayes Rule, law of total probability, etc.
- Structural properties, e.g., Independence, conditional independence
- Maximum Likelihood Estimation

# From Events to Random Variable (RV)

- Concise way of specifying attributes of outcomes
- Modeling students (Grade and Intelligence):
  - $O$  = all possible students (sample space)
  - What are events (subset of sample space)
    - Grade\_A = all students with grade A
    - Grade\_B = all students with grade B
    - HardWorking\_Yes = ... who works hard
  - Very cumbersome
- Need “functions” that maps from  $O$  to an attribute space  $T$ .
- $P(H = \text{YES}) = P(\{\text{student} \in O : H(\text{student}) = \text{YES}\})$

# Random Variables (RV)

$P(H=Yes)$



$P(H = Yes) = P(\{ \text{all students who is working hard on the course} \})$

- “functions” that maps from  $O$  to an attribute space  $T$ .

# Notations

- $P(A)$  is shorthand for  $P(A=\text{true})$
- $P(\sim A)$  is shorthand for  $P(A=\text{false})$
- Same notation applies to other **binary** RVs:  
 $P(\text{Gender}=\text{M})$ ,  $P(\text{Gender}=\text{F})$
- Same notation applies to **multivalued** RVs:  
 $P(\text{Major}=\text{history})$ ,  $P(\text{Age}=19)$ ,  $P(Q=c)$
- Note: **upper case letters/names for *variables***, **lower case letters/names for *values***

# Discrete Random Variables

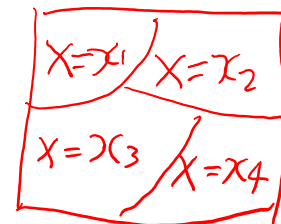
- Random variables (RVs) which may take on only a countable number of distinct values
- $X$  is a RV with arity  $k$  if it can take on exactly one value out of  $\{x_1, \dots, x_k\}$



# Probability of Discrete RV

- Probability mass function (pmf):  $P(X = x_i)$
- Easy facts about pmf
  - $\sum_i P(X = x_i) = 1$
  - $P(X = x_i \cap X = x_j) = 0$  if  $i \neq j$
  - $P(X = x_i \cup X = x_j) = P(X = x_i) + P(X = x_j)$  if  $i \neq j$
  - $P(X = x_1 \cup X = x_2 \cup \dots \cup X = x_k) = 1$

$$\sum_{i=1}^4 P(X = x_i) = 1$$



## e.g. Coin Flips

- You flip a coin
  - Head with probability  $p$ , e.g.  $=0.5$
- You flip a coin for  $k$ , e.g.,  $=100$  times
  - How many heads would you expect

e.g. Coin Flips cont.

- You flip a coin
  - Head with probability  $p$
  - **Binary** random variable
  - **Bernoulli trial** with success probability  $p$
- You flip a coin for  $k$  times
  - How many heads would you expect
  - **Number** of heads  $X$  is a discrete random variable
  - **Binomial distribution** with parameters  $k$  and  $p$

Binary =  $\{H, T\}$

$p(\#Heads)$

Integer  $\{1, 2, \dots, k\}$

# Discrete Random Variables

- Random variables (RVs) which may take on only a **countable** number of **distinct** values
  - E.g. the total number of heads  $X$  you get if you flip 100 coins
- $X$  is a RV with arity  $k$  if it can take on exactly one value out of
  - E.g. the possible values that  $X$  can take on are 0, 1, 2,..., 100

$$\{x_1, \dots, x_k\}$$

## e.g., two Common Distributions

- Uniform

- X takes values 1, 2, ...,  $N$

$$X \sim U[1, \dots, N]$$

- E.g. picking balls of different colors from a box

$$P(X = i) = 1/N$$

- Binomial

- X takes values 0, 1, ...,  $k$

$$X \sim \text{Bin}(k, p)$$

- E.g. coin flips  $k$  times

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

↓  
~ wants out k

# Today : Probability Review

- The big picture
- Events and Event spaces
- Random variables
- Joint probability, Marginalization, conditioning, chain rule, Bayes Rule, law of total probability, etc.
- Structural properties
  - Independence, conditional independence

If hard to directly estimate from data, most likely we can estimate

- 1. Joint probability
  - Use Chain Rule

$$P(A, B) = P(B) P(A|B)$$

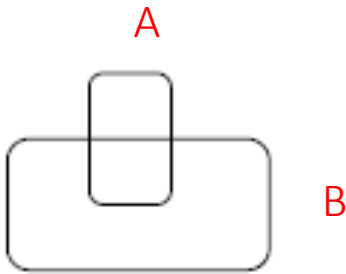
- 2. Marginal probability
  - Use the total law of probability

$$P(B) = P(B, A) + P(B, \sim A)$$
$$\parallel$$
$$P(B, A \cup \sim A) //$$

- 3. Conditional probability
  - Use the Bayes Rule

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

## Conditional / Joint / Marginal Probability



$$P(A \text{ given } B) = P(A \text{ and } B) / P(B)$$

*That is, in the frequentist interpretation, we calculate the ratio of the number of times both A and B occurred and divide it by the number of times B occurred.*

*For short we write:  $P(A|B) = P(AB)/P(B)$ ; or  $P(AB) = P(A|B)P(B)$ , where  $P(A|B)$  is the **conditional** probability,  $P(AB)$  is the **joint**, and  $P(B)$  is the **marginal**.*

*If we have more events, we use the chain rule:*

*Chain rule*

$$\underline{P(ABC)} = \underline{P(A|BC)} \underline{P(B|C)} \underline{P(C)}$$

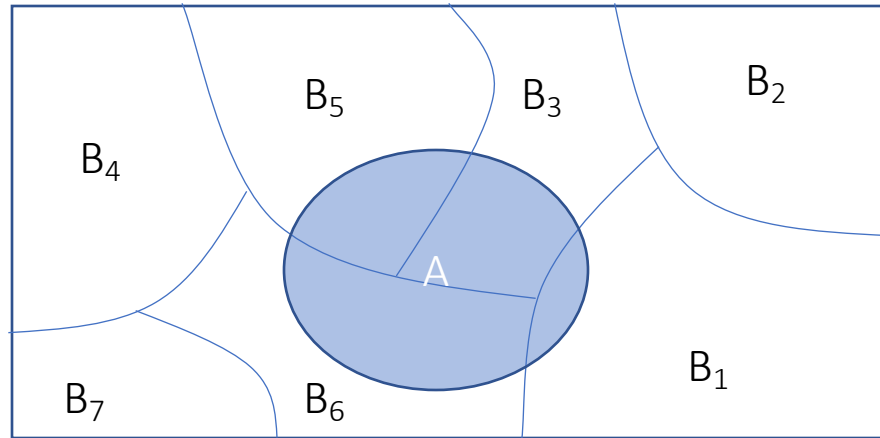


# (1). To calculate Joint Probability: Use Chain Rule

- Two ways to use chain rules on joint probability

$$\begin{array}{l} \nearrow \text{joint} \quad \nearrow \text{conditional} \\ P(A,B) = p(B|A)p(A) \rightarrow \text{marginal} \\ P(A,B) = p(A|B)p(B) \end{array}$$

(2). To calculate **Marginal Probability**:  
 Use Rule of total probability (e.g. event version)



$$P(A) = P(A \cap B) + P(A \cap \sim B)$$

$$P(B_i \cap A)$$

$$\Rightarrow p(A) = \sum P(B_i) P(A|B_i)$$

WHY ???

$$P(A) = P(A \cap \Omega)$$

$$= P(A \cap (B_1 \cup B_2 \dots \cup B_k))$$

$$= \sum P(A \cap B_i)$$

(2). To calculate **Marginal Probability**:  
**Use Rule of total probability (e.g. RV version)**

- Given two discrete RVs  $X$  and  $Y$ , which take values in:

$$\{x_1, \dots, x_k\} \quad \{y_1, \dots, y_m\}$$

$$\begin{aligned} P(X = x_i) &= \sum_j P(X = x_i \cap Y = y_j) \\ &= \sum_j P(X = x_i | Y = y_j) P(Y = y_j) \end{aligned}$$



$$P(A) = P(A \wedge B) + P(A \wedge \sim B)$$

(3). To calculate Conditional Probability:  
Use Bayes Rule (e.g. RV version)

$$P(X = x | Y = y) = \frac{P(X = x \cap Y = y)}{P(Y = y)}$$

# One Example

Assume we have a dark box with 3 red balls and 1 blue ball. That is, we have the set  $\{r, r, r, b\}$ . What is the probability of drawing 2 red balls in the first 2 tries?

$$P(B_1 = r, B_2 = r) = \underbrace{P(B_1 = r)}_{\frac{3}{4}} P(B_2 = r | B_1 = r) = \frac{1}{2}$$

$$P(B_2 = r) = P(B_1 = r, B_2 = r) + P(B_1 = b, B_2 = r)$$

$$P(B_1 = r | B_2 = r) = \frac{P(B_1 = r, B_2 = r)}{P(B_2 = r)}$$

## One Example: Joint

*Assume we have a dark box with 3 red balls and 1 blue ball. That is, we have the set  $\{\textcolor{red}{r}, \textcolor{red}{r}, \textcolor{red}{r}, \textcolor{blue}{b}\}$ . What is the probability of drawing 2 red balls in the first 2 tries?*

$$P(B_1 = r, B_2 = r) =$$

## One Example: Joint

Assume we have a dark box with 3 red balls and 1 blue ball. That is, we have the set  $\{r, r, r, b\}$ . What is the probability of drawing 2 red balls in the first 2 tries?

$$P(B_1 = r, B_2 = r) = P(B_1 = r) \underbrace{P(B_2 = r \mid B_1 = r)}_{\substack{2 \\ \downarrow \\ 3}}$$
$$P(B_1 = r) = \frac{3}{4}$$
$$P(B_1 = b) = \frac{1}{4}$$

## One Example: Joint

Assume we have a dark box with 3 red balls and 1 blue ball. That is, we have the set  $\{r, r, r, b\}$ . What is the probability of drawing 2 red balls in the first 2 tries?

$$\begin{aligned} P(B_1 = r, B_2 = r) &= P(B_1 = r) P(B_2 = r \mid B_1 = r) \\ &= \frac{3}{4} \times \frac{2}{3} = \frac{1}{2} \end{aligned}$$



## One Example: Marginal

*What is the probability that the 2<sup>nd</sup> ball drawn from the set  $\{\mathbf{r}, \mathbf{r}, \mathbf{r}, \mathbf{b}\}$  will be red?*

*Using marginalization,  $P(\mathbf{B}_2 = \mathbf{r}) = P(\mathbf{B}_2 = \mathbf{r}, \mathbf{B}_1 = \mathbf{r})$   
 $+ P(\mathbf{B}_2 = \mathbf{r}, \mathbf{B}_1 = \mathbf{b})$*

## One Example: Marginal

What is the probability that the 2<sup>nd</sup> ball drawn from the set  $\{\mathbf{r}, \mathbf{r}, \mathbf{r}, \mathbf{b}\}$  will be red?

$$\begin{aligned} \text{Using marginalization, } P(\mathbf{B}_2 = \mathbf{r}) &= P(\mathbf{B}_2 = \mathbf{r} \wedge \mathbf{B}_1 = \mathbf{r}) \\ &\quad + P(\mathbf{B}_2 = \mathbf{r} \wedge \mathbf{B}_1 = \mathbf{b}) \\ &= P(\mathbf{B}_1 = \mathbf{r}) P(\mathbf{B}_2 = \mathbf{r} | \mathbf{B}_1 = \mathbf{r}) + P(\mathbf{B}_1 = \mathbf{b}) P(\mathbf{B}_2 = \mathbf{r} | \mathbf{B}_1 = \mathbf{b}) \\ &= \frac{3}{4} \times \frac{2}{3} + \frac{1}{4} \times 1 \end{aligned}$$

One Example: Conditional

} Chain Rule  
total law Prob

$$P(B_1=r | B_2=r) = \frac{P(B_2=r | B_1=r) P(B_1=r)}{P(B_2=r)}$$

⇒ lost last page

⇓ Last

$$= \frac{P(B_2=r | B_1=r) P(B_1=r)}{P(B_2=r, B_1=r) + P(B_2=r, B_1=b)}$$

# Bayes Rule

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

This is Bayes Rule

**Bayes, Thomas (1763)** An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, **53:370-418**



$$\frac{P(Y=\text{yes}) P(X|Y=\text{yes})}{P(X)}$$

if  $P(Y=\text{yes}|X) > P(Y=\text{No}|X)$

$$\frac{P(Y=\text{No}) P(X|Y=\text{No})}{P(X)}$$

$\Rightarrow \hat{y} = \text{yes}$

# More General Forms of Bayes Rule

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

$P(B_2=t, B_1=t)$   
 $P(B_2=t, B_1=t) + P(B_2=t, B_1=b)$

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

$$P(A = a_1 | B) = \frac{P(B | A = a_1)P(A = a_1)}{\sum_i P(B | A = a_i)P(A = a_i)}$$

E.g.: Use both Bayes Rule and Marginal

- X and Y are discrete RVs...

$$P(X = x_i | Y = y_j) = \frac{P(X = x_i \cap Y = y_j)}{P(Y = y_j)}$$

$$\{x_1, \dots, x_k\} \downarrow$$

$$P(X = x_i | Y = y_j) = \frac{P(Y = y_j | X = x_i) P(X = x_i)}{\sum_k P(Y = y_j | X = x_k) P(X = x_k)}$$

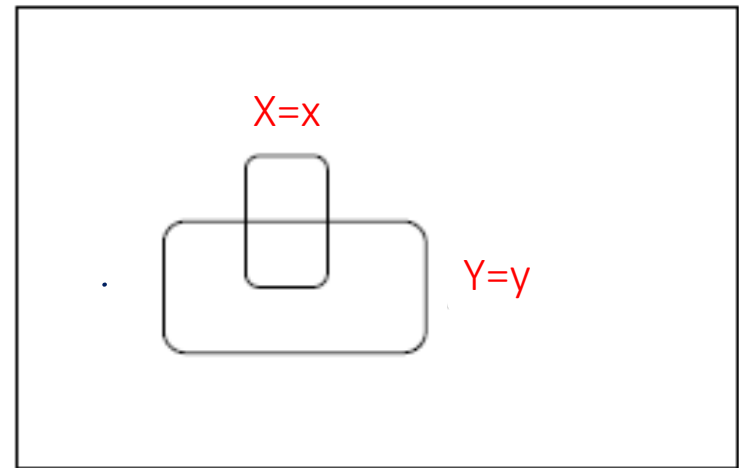
# Simplify Notation: Conditional Probability

$$P(X = x | Y = y) = \frac{P(\text{X} = x \cap \text{Y} = y)}{P(\text{Y} = y)}$$

events

But we will always write it this way:

$$P(x | y) = \frac{p(x, y)}{p(y)}$$



$P(X=x \text{ true}) \rightarrow P(X=x) \rightarrow P(x)$

$P(x) \leftarrow P(\underline{X}=x) \leftarrow P(\underline{X}=x \text{ true})$   
 value                      RV                      event

# Simplify Notation:

## An Example of estimating conditional

- We know that  $P(\text{rain}) = 0.5$
- If we also know that the grass is wet, then how this affects our belief about whether it rains or not?

$$P(\text{rain} \mid \text{wet}) = \frac{P(\text{rain})P(\text{wet} \mid \text{rain})}{P(\text{wet})}$$

$W =$       $G =$



# Simplify Notation:

## An Example of estimating conditional

- We know that  $P(\text{rain}) = 0.5$
- If we also know that the grass is wet, then how this affects our belief about whether it rains or not?

$$\overset{W=}{P}(\overset{G=}{rain} | wet) = \frac{P(rain)P(wet | rain)}{P(wet)}$$

$P(W=S | wet)$

$$P(x | y) = \frac{P(x)P(y | x)}{P(y)} = \frac{p(x, y)}{p(y)}$$

# Simplify Notation: Conditional

- Bayes Rule

$$P(x | y) = \frac{P(x)P(y | x)}{P(y)}$$

- You can condition on **more variables**

$$P(x | y, z) = \frac{P(x | z)P(y | x, z)}{P(y | z)}$$

# Simplify Notation: Marginal

- We know  $p(X, Y)$ , what is  $P(Y=y)$  or  $P(X=x)$ ?
- We can use the law of total probability

$$p(x) = \sum_y P(x, y)$$
$$= \sum_y P(y)P(x|y)$$

$\{y_1, \dots, y_m\}$

all possible  $Y$  values

$$p(x) = \sum_{y,z} P(x, y, z)$$
$$= \sum_{z,y} P(y, z)P(x|y, z)$$

$\sum_y \sum_z p(y, z) = 1$

# Simplify Notation:

## An Example

- We know that  $P(\text{rain}) = 0.5$
- If we also know that the grass is wet, then how this affects our belief about whether it rains or not?

$$P(\text{rain} \mid \text{wet}) = \frac{P(\text{rain})P(\text{wet} \mid \text{rain})}{P(\text{wet})}$$

Handwritten annotations:

- $P(\text{rain})$  is annotated with  $0.5$ .
- $P(\text{wet} \mid \text{rain})$  is annotated with  $1$ .
- $P(\text{wet})$  is annotated with  $P(\text{wet, rain}) + P(\text{wet, sunny})$ .
- The denominator is expanded to  $P(\text{rain})P(\text{wet} \mid \text{rain}) + P(\text{sunny})P(\text{wet} \mid \text{sunny})$ .
- Below the equation, a large blue bracket groups the terms  $P(\text{rain})P(\text{wet} \mid \text{rain})$  and  $P(\text{sunny})P(\text{wet} \mid \text{sunny})$ .
- Below the bracket, the terms are labeled: "Weather {rain, sunny}" and "Grass {wet, dry}".
- The term  $P(\text{wet} \mid \text{sunny})$  is underlined in red.

# Today : Probability Review

- The big picture
- Events and Event spaces
- Random variables
- Joint probability, Marginalization, conditioning, chain rule, Bayes Rule, law of total probability, etc.
- Structural properties, e.g., Independence, conditional independence
- Maximum Likelihood Estimation

# Independent RVs

- Definition:  $X$  and  $Y$  are independent *iff*

$$P(X = x \cap Y = y) = P(X = x)P(Y = y)$$

## More on Independence

$$P(X = x \cap Y = y) = P(X = x)P(Y = y)$$



$$P(X = x | Y = y) = P(X = x)$$



$$P(Y = y | X = x) = P(Y = y)$$

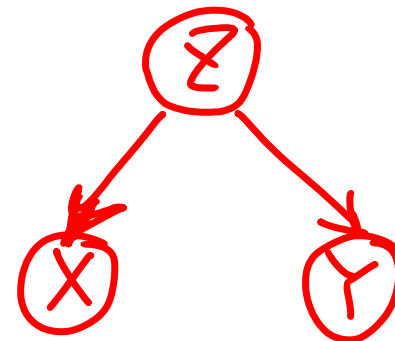
- E.g. no matter how many heads you get, your friend will not be affected, and vice versa

# More on Independence

- $X$  is independent of  $Y$  means that knowing  $Y$  does not change our belief about  $X$ .
- The following forms are **equivalent**:
  - $P(X=x, Y=y) = P(X=x) P(Y=y)$
  - $P(X=x | Y=y) = P(X=x)$
- The above should hold for all  $x_i, y_j$
- It is symmetric and **written as**  $X \perp Y$



# Conditionally Independent RVs

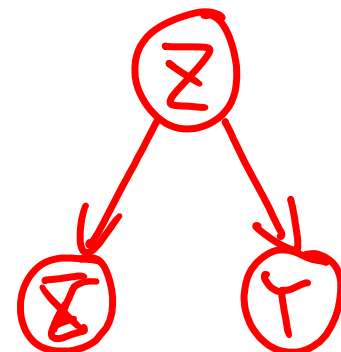


- Intuition: X and Y are conditionally independent given Z means that once Z is **known**, the value of X does not add any **additional** information about Y
- Definition: X and Y are conditionally independent given Z *iff*

$$P(X = x \cap Y = y | Z = z) = P(X = x | Z = z) P(Y = y | Z = z)$$

If holding for all  $x_i, y_j, z_k$

$$X \perp Y | Z$$



## More on Conditional Independence

$$P(X = x \cap Y = y | Z = z) = P(X = x | Z = z) P(Y = y | Z = z)$$



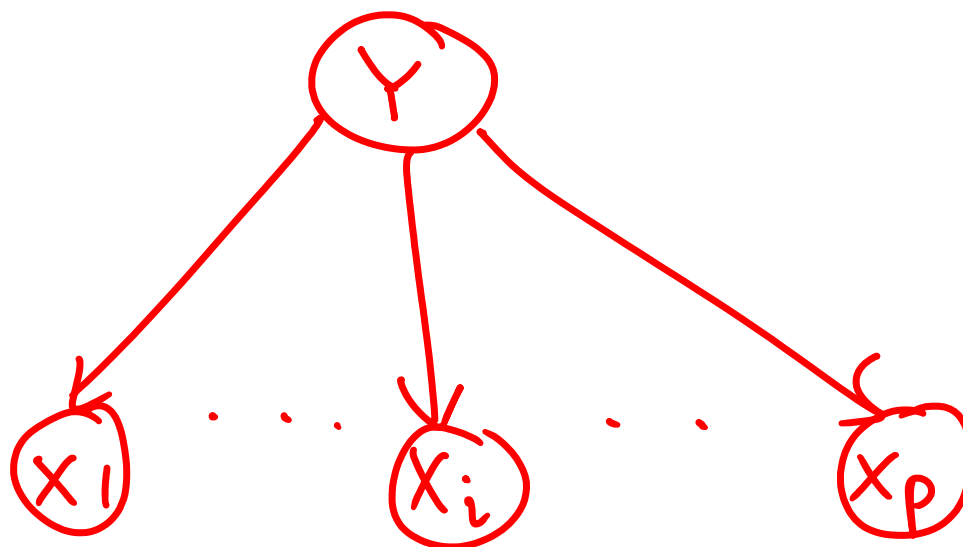
$$P(X = x | Y = y, Z = z) = P(X = x | Z = z)$$



$$P(Y = y | X = x, Z = z) = P(Y = y | Z = z)$$

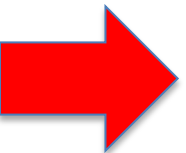
# independence and conditional independence

- Independence does not imply conditional independence.
- Conditional independence does not imply independence.



# Today Recap: Probability Review

- The big picture
- Events and Event spaces
- Random variables
- Joint probability, Marginalization, conditioning, chain rule, Bayes Rule, law of total probability, etc.
- Structural properties, e.g., Independence, conditional independence
- Maximum Likelihood Estimation (next class)



# References

- Prof. Andrew Moore's review tutorial
- Prof. Nando de Freitas's review slides
- Prof. Carlos Guestrin recitation slides